Beijing 2002
August 20–28

Proceedings of the

International Congress of Mathematicians

Vol. I : Plenary Lectures and Ceremonies

Higher Education Press

International Congress of Mathematicians (2002, Beijing)
Proceedings of the International Congress of Mathematicians
August 20–28, 2002, Beijing

Editor: LI Tatsien (LI Daqian)
Department of Mathematics, Fudan University
Shanghai 200433, China Email: dqli@fudan.edu.cn

Editorial Assistants: Cai Zhijie, Xue Mi, Zhou Chunlian

This Volume contains information on the organization of the Congress including a list of the participants, the speeches at the opening and closing ceremonies, and the reports on the work of the Fields Medalists and the Nevanlinna Prize Winner as the first part, and the Plenary Lectures together with some Invited Lectures which have not been included in Volumes II and III as the second part.

The electronic version of Volumes I, II and III will be freely available on the web as a Math ArXiv overlay at the web page

http://front.math.ucdavis.edu/ICM2002

Copy Editors: Li Rui, Zhang Xiaoping

ISBN 7-04-008690-5 Set of 3 Volumes

Contents

Volume I

The Work of the Fields Medalists and of the Rolf Nevanlinna Prize Winner

Invited One-Hour Plenary Lectures

Appendix A: Invited Forty-Five Minute Lectures at the Section Meetings (not included in Volumes II and III)

Volume II

Section 1. Logic

Section 2. Algebra

Section 3. Number Theory

Section 4. Differential Geometry

Section 5. Topology

Section 6. Algebraic and Complex Geometry

Section 7. Lie Group and Representation Theory

Section 8. Real and Complex Analysis

Section 9. Operator Algebras and Functional Analysis

Volume III

Section 10. Probability and Statistics

Section 11. Partial Differential Equations

Section 12. Ordinary Differential Equations and Dynamical Systems

Section 13. Mathematical Physics

Section 14. Combinatorics

Section 15. Mathematical Aspects of Computer Science

Section 16. Numerical Analysis and Scientific Computing

Section 17. Application of Mathematics in the Sciences

Section 18. Mathematics Education and Popularization of Mathematics

Section 19. History of Mathematics

Preface

The Proceedings of the International Congress of Mathematicians 2002 (ICM2002), held in Beijing from August 20 to August 28, are published in three volumes.

Volume I contains information on the organization of the Congress including a list of the participants, the speeches at the opening and closing ceremonies, and the reports on the work of the Fields Medalists and the Nevanlinna Prize Winner as the first part, and the Plenary Lectures together with some Invited Lectures which have not been included in Volumes II and III as the second part. Volumes II and III which contain most of the Invited Lectures were published and distributed to the participants at the beginning of the Congress. Since the first part of material must be gathered during or after the Congress, Volume I is published several months later.

The electronic version of Volumes I, II and III will be freely available on the web as a Math ArXiv overlay at the web page

`http://front.math.ucdavis.edu/ICM2002`

We take this opportunity to express our sincere thanks to all the speakers and organizers for their contribution and cooperation.

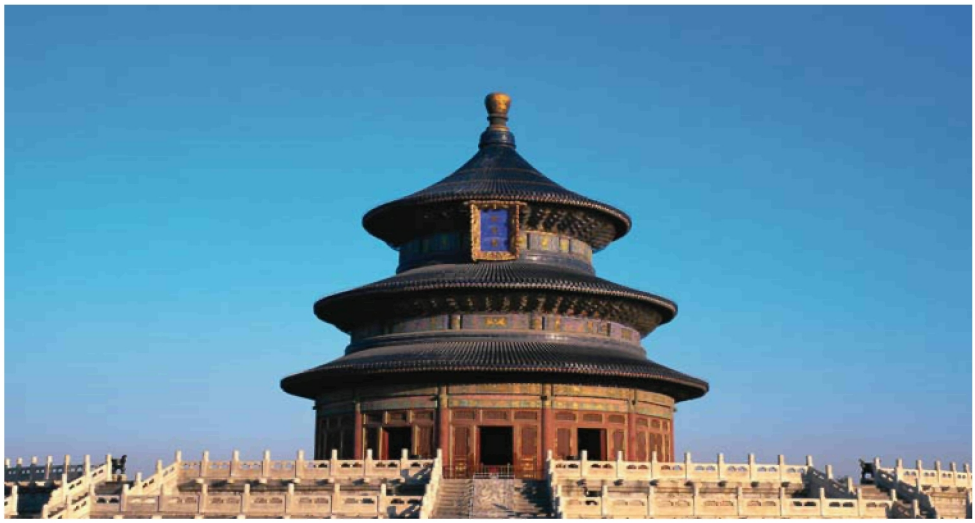
We are also very grateful to Higher Education Press for the publication of the Proceedings of ICM2002.

Shanghai, November 2002

LI Tatsien (LI Daqian)
Fudan University
Shanghai 200433, China
Email: dqli@fudan.edu.cn

Past Congresses

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



2002 Beijing

Past Fields Medalists and Rolf Nevanlinna Prize Winners

Fields Medalists

1936 Lars V. Ahlfors
Jesse Douglas

1950 Laurent Schwartz
Atle Selberg

1954 Kunihiko Kodaira
Jean-Pierre Serre

1958 Klaus F. Roth
René Thom

1962 Lars Hörmander
John W. Milnor

1966 Michael F. Atiyah
Paul J. Cohen
Alexander Grothendieck
Steve Smale

1970 Alan Baker
Heisuke Hironaka
Sergei P. Novikov
John G. Thompson

1974 Enrico Bombieri
David B. Mumford

1978 Pierre R. Deligne
Charles F. Fefferman
Grigorii A. Margulis
Daniel G. Quillen

1982 Alain Connes
William P. Thurston
Shing-Tung Yau

1986 Simon K. Donaldson
Gerd Faltings
Michael H. Freedman

1990 Vladimir G. Drinfeld
Vaughan F. R. Jones
Shigefumi Mori
Edward Witten

1994 Jean Bourgain
Pierre-Louis Lions
Jean-Christophe Yoccoz
Efim Zelmanov

1998 Richard E. Borcherds
William T. Gowers
Maxim Kontsevich
Curtis T. McMullen
Andrew Wiles (special tribute)

Rolf Nevanlinna Prize Winners

1982 Robert E. Tarjan
1986 Leslie G. Valiant
1990 Alexander A. Razborov
1994 Avi Wigderson
1998 Peter W. Shor

The Committees of the Congress

Program Committee

Yuri Manin	Chair, Max-Planck-Institut fur Math., Bonn
John Ball	University of Oxford
Franco Brezzi	University of Pavia
Gérard Laumon	Univ. de Paris-Sud
Laszlo Lovasz	Yale University
Tetsuji Miwa	Kyoto University
Peter Sarnak	Princeton University
Alain-Sol Sznitman	ETH-Zentrum, Zurich
Gang Tian	Massachusetts Institute of Technology
Michèle Vergne	Ecole Normale Supérieure, Paris
Wen-Tsun Wu	Academia Sinica, Beijing

Panels ICM2002

1. Logic

Ehud Hrushovski	Chair, ehud@math.huji.ac.il
A. Wilkie	wilkie@maths.ox.ac.uk
Yuri Matiyasevich	yumat@pdmi.ras.ru
A. Louveau	louveau@ccr.jussieu.fr
Donald A. Martin	dam@math.ucla.edu

2. Algebra

David Eisenbud	Chair, de@msri.org
Eric Friedlander	eric@math.nwu.edu
De Concini	deconcin@mat.uniroma1.it
Michael Aschbacher	asch@its.caltech.edu
Michel Van den Bergh	vdbergh@luc.ac.be
Raman Parimala	parimala@math.tifr.res.in
Idun Reiten	idunr@math.ntnu.no
Aner Shalev	shalev@math.huji.ac.il
Andrei Zelevinsky	andrei@neu.edu

3. Number Theory

Don Zagier	Chair, don@mpim-bonn.mpg.de
Barry Mazur	mazur@math.harvard.edu
Colliot–Thélène	colliot@math.tifr.res.in

8. Real and Complex Analysis

Michael Christ	Chair, mchrist@math.berkeley.edu
(appointed after the death of Thomas Wolff, initial Chair)	
Duong H. Phong	dp@math.columbia.edu
Jean-Pierre Demailly	Jean-Pierre.Demailly@ujf-grenoble.fr
Seppo Rickman	rickman@cc.helsinki.fi
Kari Astala	astala@math.lsa.umich.edu
	astala@math.jyu.fi
Jean-Michel Bony	bony@math.polytechnique.fr
Peter Jones	jones@math.yale.edu

9. Operator Algebras and Functional Analysis

Dan Voiculescu	Chair, dvv@math.berkeley.edu
Henri Moscovici	henri@math.ohio-state.edu
W.T. Gowers	W.T.Gowers@dpmms.cam.ac.uk
Vaughan Jones	vfr@math.berkeley.edu
Gilles Pisier	gip@ccr.jussieu.fr
Georges Skandalis	skandal@math.jussieu.fr

10. Probability and Statistics

Jean-Francois Le Gall	Chair, Jean-Francois.Le.Gall@ens.fr
David Donoho	donoho@stat.stanford.edu
Mark Freidlin	mif@math.umd.edu
Charles Newman	newman@courant.nyu.edu
David Aldous	aldous@stat.berkeley.edu
Friedrich Goetze	goetze@mathematik.Uni-Bielefeld.DE
Ildar Ibragimov	ibr32@pdmi.ras.ru
Shinichi Kotani	kotani@math.sci.osaka-u.ac.jp
Zhi-Ming Ma	mazm@amath8.amt.ac.cn

11. Partial Differential Equations

Neil Trudinger	Chair, neil.trudinger@anu.edu.au
Constantine M. Dafermos	dafermos@cfm.brown.edu
Fang-Hua Lin	linf@math1.cims.nyu.edu
Ding Wei-yue	dingwy@public.bta.net.cn
Richard Melrose	rbm@math.mit.edu
Mariano Giaquinta	giaquinta@sns.it
Olli Martio	olli.martio@helsinki.fi

12. ODE and Dynamical Systems

Marcelo Viana	Chair, viana@impa.br
John Norman Mather	jnm@math.princeton.edu
S. Kuksin	s.b.kuksin@ma.hw.ac.uk
D. Anosov	anosov@mi.ras.ru
Etienne Ghys	ghys@umpa.ens-lyon.fr
Michael Jakobson	mvy@math.umd.edu
Harry Furstenberg	harry@math.huji.ac.il

13. Mathematical Physics

Robbert Dijkgraaf	Chair, rhd@fwi.uva.nl
Isadore Singer	ims@math.mit.edu
Fedor Smirnov	smirnov@lpthe.jussieu.fr
M. Aizenman	aizenman@princeton.edu
M. Jimbo	jimbo@kusm.kyoto-u.ac.jp
D. Ruelle	ruelle@ihes.fr
A. Zamolodchikov	sashaz@physics.rutgers.edu

14. Combinatorics

Paul D. Seymour	Chair, pds@math.princeton.edu
Bernd Sturmfels	bernd@math.berkeley.edu
G. Kalai	kalai@math.huji.ac.il
Zoltan Furedi	z-furedi@math.uiuc.edu
Colin McDiarmid	cmcd@stats.ox.ac.uk

15. Mathematical Asoects of Computer Science

A. Wigderson	Chair, avi@math.ias.edu
A. Schrijwer	lex@cwi.nl
Mark Jerrum	mrj@dcs.ed.ac.uk
Amir Pnueli	amir@wisdom.weizmann.ac.il
Oded Goldreich	oded@wisdom.weizmann.ac.il

16. Numerical Analysis and Scientific Computing

Roland Glowinski	Chair, roland@math.uh.edu
Gilbert Strang	gs@math.mit.edu
W. Hackbusch	wh@mis.mpg.de
Lin Qun	qlin@staff.iss.ac.cn
A.Bjorck	akbj@mai.lin.se
R.Jeltsch	jeltsch@math.ethz.ch
I.Sloan	sloan@maths.unsw.edu.au
N.Trefethen	lnt@comlab.ox.ac.uk

17. Applications of Mathematics in the Sciences

Robert V. Kohn	kohn@cims.nyu.edu
Jerry E. Marsden	marsden@cds.caltech.edu
Etienne Pardoux	pardoux@gyptis.univ-mrs.fr
Bjorn Engquist	engquist@math.ucla.edu
Keith Glover	kg@eng.cam.ac.uk
Frank Hoppensteadt	fchoppen@asu.edu
Joe Keller	keller@math.stanford.edu
Pierre-Louis Lions	lions@dmi.ens.fr
Stephane Mallat	mallat@cmapx.polytechnique.fr
Linda Petzold	petzold@engineering.ucsb.edu

18. Mathematics Education and Popularization of Mathematics

Hyman Bass	Chair, hybass@umich.edu
Bernard Hodgson	bhodgson@mat.ulaval.ca
B. J. Jiang	jiangbj@sxx0.math.pku.edu.cn
Michele Artigue	France
Jeremy Kilpatrick	USA
Mogens Niss	Denmark
Miguel de Guzman	Spain
Lee Peng Yee	Singapore
Igor Sharygin	Russia
Gilah Leder	Australia

19. History of Mathematics

Henk Bos	Chair, bos@math.uu.nl
Jeremy Gray	j.j.gray@open.ac.uk
Li Wenlin	wli@math08.math.ac.cn
S. Demidov	serd@ssd.pvt.msu.su
Kirsti Andersen	ivhka@ifa.au.dk
Michio Yano	yanom@cc.kyoto-su.ac.jap

Fields Medal Committee

Yakov Sinai	Chair, Princeton University
James Arthur	University of Toronto
Spencer Bloch	University of Chicago
Jean Bourgain	Institute for Advanced Study
Helmut Hofer	Courant Institute, New York
Yasutaka Ihara	Kyoto University
H. Blaine Lawson	SUNY Stony Brook
Sergei Novikov	University of Maryland
George Papanicolaou	Stanford University
Efim Zelmanov	Yale University

Nevanlinna Committee

Michael Rabin	Chair, Harvard University
Andrei Agrachev	Steklov Institute
Ingrid Daubechies	Princeton University
Wolfgang Hackbusch	University of Kiel
Alexander Schrijver	CWI, Amsterdam

Honorary President of ICM2002

Shing-Shen Chern

President of ICM2002

Wu, Wen-Tsun

Honorary Committee of ICM2002

Ding, Shisun
Hu, Guoding
Gu, Chaohao
Ke, Zhao
Li, Zhengdao
Su, Buqing
Wang, Yuan
Wu, Wen-Tsun
Yang, Zhenning
Zhou, Guangzhao
Zhu, Guangya

Steering Committee

Chen, Jiaer
Chen, Liangyu
Chen, Zhili
Liu, Qi
Lu, Yongxiang
Song, Jian
Xiang, Huaicheng
Xu, Guanhua
Xu, Kuangdi
Zhang, Yutai
Zhao, Qizheng
Zhu, Lilan

Shing-Shen Chern (left) and John F. Nash, Jr.

Organizing Committee

Chang, Kung Ching	Chair (before 2000), Peking Univ.
Ma, Zhiming	Chair (after 2000), AMSS, CAS
Chen, Shuping	Zhejiang Univ. and Guizhou Univ.
Ding, Weiyue	AMSS, CAS and Peking Univ.
Feng, Keqin	Tsinghua Univ.
Feng, Qi	AMSS, CAS
Hou, Zixin	Nankai Univ.
Jiang, Boju	Peking Univ.
Li, Daqian (Li, Tatsien)	Fudan Univ.
Li, Wenlin	AMSS, CAS
Lin, Qun	AMSS, CAS
Lin, Fanghua	Univ. of New York
Liu, Taiping	Inst. of Math., Academia Sinica
Liu, Yingming	Sichuan Univ.
Lu, Shanzhen	Beijing Normal Univ.
Wang, Jianpan	East China Normal Univ.
Wong, Roderrick	City Univ. of Hong Kong
Yang, Lo	AMSS, CAS
Yuan, Ya-xiang	AMSS, CAS
Zhang, Jiping	Peking Univ.
Zhang, Xiangsun	AMSS, CAS
Zhou, Qing	East China Normal Univ.

Local Scientific Committee

Ding, Weiyue	Chair, AMSS, CAS and Peking Univ.
Yan, Jia-an	Vice Chair, AMSS, CAS
Feng, Qi	Vice Chair, AMSS, CAS
Fan, Genghua	AMSS, CAS
Feng, Keqin	Tsinghua Univ.
Gao, Xiaoshan	AMSS, CAS
Li, Bingren	AMSS, CAS
Li, Jiayu	AMSS, CAS and Fudan Univ.
Li, Wenlin	AMSS, CAS
Shen, Longjun	Beijing Inst. of Applied Physics and Comp. Math.
Sun, Xiaotao	AMSS, CAS
Wang, Shi-cheng	Peking Univ.
Wang, Shi-kun	AMSS, CAS
Wen, Lan	Peking Univ.
Xiao, Jie	Tsinghua Univ.
Xi, Nanhua	AMSS, CAS
Ye, Qixiao	Beijing Inst. of Tech.
Zhou, Xianyu	AMSS, CAS

Other Subcommittees Appointed by the Organizing Committee

Except the Local Scientific Committee, the Organizing Committee has also set various other sub-committees. These sub-committees and their presidents are as follows.

Liaison to IMU:	Chang, Kung Ching	Peking Univ.
Publication Sub-committee:	Li, Daqian (Li, Tatsien)	Fudan Univ.
Financial Sub-committee:	Lin, Qun	AMSS, CAS
Fund-raising Sub-committee:	Hou, Zixin	Nankai Univ.
Satellite Conference Sub-committee:	Li, Wenlin	AMSS, CAS
Grant Sub-committee:	Lu, Shanzhen	Beijing Normal Univ.
Local Arrangement Sub-committee:	Zhang, Jiping	Peking Univ.
Network and Website Maintenance Sub-committee:	Jin, Yafeng	AMSS, CAS

Secretariat Office

Yuan, Ya-xiang	AMSS, CAS, General Secretary
Feng, Qi	AMSS, CAS, Associate General Secretary
Peng, Lizhong	Peking Univ., Associate General Secretary
Guo, Wei	Hebei Univ. of Technology
Li, Dong	AMSS, CAS
Li, Juan	Inst. of Chemistry, CAS
Liu, Feng	BICC
Wu, Jinrong	AMSS, CAS
Zhang, Jinyu	AMSS, CAS

P. Griffths, J. Palis, D. Mumford

List of Sponsors, Donors and Contributors

The Organizing Committee is greatly indebted to various organizations for their financial supports and other helps. The Congress would have not been possible without the following sponsorship:

Ministry of Finance of PRC
Ministry of Science and Technology of PRC
Ministry of Education of PRC
Chinese Academy of Sciences
National Natural Science Foundation of China
Beijing Municipal Government
China Association for Science and Technology

The Organizing Committee would like to thank many institutions, corporations and individuals for their generous donations and/or contributions, particularly to:

K.C. Wong Education Foundation
Xing An Securities Co. Ltd.
Academy of Mathematics and System Sciences
Shing-Shen Chern
School of Mathematical Science, Peking University
School of Math. and Inst. of Math., Nankai University
Dept. of Math. and Inst. of Math., Fudan University
Dept. of Math. Sci., Tsinghua University
Dept. of Math., Beijing Normal University
Shanghai Jiaotong University
School of Math. & Computer Sci., Nanjing Normal University
Dept. of Math., East China Normal University
School of Sciences, Tongji University
Merry Sun Investment Co. Ltd.
Lu Neng Inform. Tech. Co. Ltd. of Shandong University
Di Wei Software Company of Shandong University
Shanghai Fudan Science Park Co. Ltd.
Liu Bie Ju Center for Math. Sci., City University of Hong Kong
Dept. of Math., Capital Normal University
School of Math., Sichuan University
Shanghai University
Math. and Stat. School, Wuhan University
Fan Wei
Shenzhen Shenzhan Bus Co. Ltd. Arbiter Li An Sheng
Shanghai Hua-Shen Sino-Foreign Cultural Exchange Service Co. Ltd.

ICM 2002 · Vol. I · 13–26

Opening Ceremony

The opening ceremony of the Congress was held at the Great Hall of People on Tuesday, August 20, 2002, staring at 3:00pm. Jiang Zemin, President of the People's Republic of China, was present and granted the Fields Medals to two Fields Medalists. Here are the speeches at the opening ceremony.

Jacob Palis

President of the International Mathematical Union

Dear Colleagues, Ladies and Gentleman,

I am greatly honored and pleased to welcome you all to ICM2002, the 24th International Congress of Mathematicians.

This is in many ways a very special Congress. Indeed, it is the first in the new Millennium and, therefore, we are bringing the dreams of Cantor and Felix Klein, dreamed in the late 1900s, into the 21st Century. They realized, then, that mathematics was becoming too large and diversified a subject and that was almost impossible for one person to embrace, like probably was the case of Monge, Laplace, Lagrange and Gauss, among others, at the turn of the 19th Century. Thus, interaction among mathematicians both at a national and international level was the clear road for its development. Their dream was not only robust in time, but has grown in dimension; mathematics has become more and more international, and solidarity across countries has been increasing at a fast pace. This is occurring not only at a world basis, particularly through the activities of IMU, among which the ICM is a major event, but also in regional scenarios, as indicated by the rather recent creations of the European Mathematical Society and the Latin American and Caribbean Mathematical Union, following that of the African Mathematical Union and of the International Council for Industrial and Applied Mathematics. The first two organizations are affiliated to IMU, and we have solid relations with the last ones.

The 24th ICM is also unique because for the first time it is taking place in a developing country, and in fact in the fastest growing country in the world at present, with a population which is about a fourth of humanity. Per se this makes the ICM more inclusive and being inclusive is a basic principle of our Union, as also shown by our joint efforts with the Local Organizing Committee in providing the opportunity to more than 400 colleagues, young and senior, from less affluent

parts of the world, to participate in the Congress. By having the Congress here, we are giving our trust to China for its commitment to mathematics and in particular to its young talents. But China is also paying a precious tribute to the Union, by the presence among us, for the first time in our history, of the highest authority of the host country, President Jiang Zemin. About a year and a half ago, he accepted our invitation to be in this Opening Ceremony and jointly with us award the Fields Medals. In doing so, the President is showing his appreciation for our science and its importance to the world of today. We are very confident that the Congress here in China will mark a formidable change in the level and scope of activities of mathematics in this country: a tree that was planted by S. S. Chern, L. K. Hua and K. Feng, as well as by C. H. Gu, W. T. Wu and S. Liao, and more recently S. T. Yau and G. Tian, among others.

This Congress is also a culmination of an intense period of activities in mathematics throughout the world, as well as for achieving a certain maturity concerning the perspective for its future development. In this respect, besides fundamental research, the importance of the interaction of mathematics with other areas of science, beyond the classical case of physics, is now largely accepted. Also, more emphasis in applications is to be given. Moreover, there should be no division between pure and applied mathematics in accordance with Pasteur's beautiful sentence that there is no applied science, but applications of science. In terms of activities, we had an intense celebration of the Year 2000 as the World Mathematical Year: IMU pub-

The opening ceremony of the ICM2002

lished a book "Mathematics: Frontiers and Perspectives"; co-sponsored major conferences in Europe, Latin America, Africa and Asia, one of them through its Commission on Mathematical Instruction, and promoted many mathematical exhibitions and events directed to the general public. Such a celebration was part of a Declaration made by Jacques-Louis Lions, in Rio de Janeiro, in 1992.

Unfortunately, I have to register that he, Jurgen Moser and Lion's former adviser, the Fields Medallist Laurent Schwartz passed away in the last years. Of prime importance in this period, has been the activity of the Union's Committee on Electronic Information and Communication and the work of the IMU Commissions on Development and Exchanges (CDE), Instruction (ICMI) and History (ICHM).

The present Congress is also special in other ways. For the first time, the IMU General Assembly has elected a woman to its Executive Committee and also a Chinese. Furthermore, at this occasion, the mathematical community can commemorate the creation of two new prizes. The first, called the Gauss Prize for Applications of Mathematics is to be jointly awarded once every four years by IMU and the German Mathematical Society. The second, in honor of Abel, shall be awarded every year by the Norwegian Academy of Sciences: similar to the Nobel Prize, it has the potential to change, in years to come, the landscape of mathematics in the world scenario of sciences.

Finally, on behalf of all of us, I wish to express our sincere gratitude to the Chinese Institutions that made the Congress possible and most especially to our colleagues Zhiming Ma, K.C. Chang, Daqian Li, Weiyue Ding and Ya-xiang Yuan for their warm reception and excellent organization.

Thank you very much.

The scene of the opening ceremony of the ICM2002

Li Lanqing

Vice Premier of the People's Republic of China

Respected President Jiang Zemin, Respected IMU President Mr. Palis, Distinguished Guests, Ladies and Gentlemen:

Today, mathematicians from all over the world are gathering here for the first International Congress of Mathematicians in the new millennium. On behalf of President Jiang Zemin and the Chinese government, I have the pleasure to extend to you our warmest welcome.

No one could have imagined the extraordinary evolution of science and technology over the past century. Space exploration, nuclear energy, computers and information technology, not to mention biological engineering, are all milestones that mark a new era of knowledge for humankind. Our social progress depends on scientific innovation, and mathematics is fundamental to science. Mathematics expressed the theory of relativity and the quantum mechanics in the early 20th century; since then mathematicians has played a vital role in inventing computers, designing space and energy programs, and investigating the structure of DNA molecules. Mathematics is the language of the universe.

Mathematical methods are used extensively in economics, medicine, agriculture, architecture, arts and all other fields of modern knowledge. As Roger Bacon pointed out, mathematics is the key to all branches of science. Today mathematics is the keystone of high technology, and, in a sense, the symbol of modern civilization. In this light, the Chinese government is especially delighted to see this congress being held in Beijing. As President Jiang Zemin clearly expressed when he met with Professor Chern Shing-shen, IMU President Palis and other mathematicians in October 2000, "the Chinese government fully supports hosting the 2002 International Congress of Mathematicians in Beijing. China wishes to take this opportunity to promote math research and education in the country, in an effort to bring them up to the world advanced level in the early 21st century and lay a solid foundation for the future progress of science and technology in China."

As a developing country, China is marching on the road toward modernization. It has been a century-long pursuit for the Chinese people to revitalize their country through development of science and education. This historical process has been even further accelerated in the last two decades by reform and opening up policies, as both young talents and accomplished experts emerge in great numbers on the international scientific scene. The Chinese government has fully supported all endeavors to pursue this development, including a series of programs launched nationwide to

promote basic scientific research, especially in mathematics. For example, in the past four years, the National Science Foundation of China has doubled its funding for mathematics, and the government has allocated thousands of millions of yuan to support the Pilot Knowledge Innovation Program in the Chinese Academy of Sciences. We are aware that China still has a long way to go before reaching the advanced world levels in science and technology. Science knows no boundaries. The advancement of science requires peace, stability and cooperation. In this regard, I believe that the International Congresses of Mathematicians, with over a hundred years of tradition, sets the example. Hosting the 24th Congress in Beijing is a good opportunity for Chinese scientists to learn from and to cooperate with their colleagues abroad. I hope that this congress will mark a new starting point for the development of mathematics and science in China. As the first congress ever held in a developing country, I also hope that this congress will inspire a new era of international cooperation for global scientific community.

In about 10 minutes' time, the new Fields medallists and the winner of the Nevanlinna Prize will be announced and awarded. I would like to take this opportunity to offer them my sincere congratulations. Their achievements not only represent their distinguished contributions to mathematics, but to world cooperation and the well-being of all humankind.

In conclusion, I wish this congress a great success, and all our guests a memorable stay in China.

Thank you!

Beijing International Conference Center

Shing-shen Chern

Honorary President of the ICM2002

It is my great pleasure to welcome you to this gathering. We are in an ancient country that is very different from Western Europe where modern mathematics started. In 2000, we had a mathematics year, an effort to attract more people to math. We now have a vast field and a large number of professional mathematicians whose major work is mathematics. Mathematics used to be individual work. But now we have a public. In such a situation a prime duty seems to be to make our progress available to the people. There is clearly considerable room for popular expositions. I also wonder if it is possible for research articles to be produced by a historical and popular introduction. The net phenomenon could be described as a globalization. It is more than geographical. In recent studies different fields were not only found to have contacts, but were merging. We might even foresee a unification of mathematics, including both pure and applied, and even the possibility of the emergence of a new Gauss.

China has a long way to go in modern mathematics. In recent contests of the international mathematical Olympiad China has consistently done very well. Thus China has begun from the roots and China has the advantage of "number" (of people). Hopefully this Congress will be a critical point in the development of modern math in China. The great Confucius guided China spiritually for over 2000 years. The main doctrine is "仁" (pronounced "ren"), meaning two people, i.e., human relationship. Modern science has been highly competitive. I think an injection of the human element will make our subject more healthy and enjoyable. Let us wish that this Congress will open a new era in the future development of math.

Zhou Guangzhao

Vice Chairman, Standing Committee of NPC, China
President of the China Association for Science and Technology

Ladies and gentlemen,

Today, we are particularly overjoyed at the grand opening of the 24th International Congress of Mathematicians. On behalf of the China Association for Science and Technology and the Chinese scientific community, I would like to express our warmest welcome to participants from all over the world and our sincere congratulations to the newly awarded Fields medallists and the winner of Nevalinna Prize.

The reason of our being particularly overjoyed lies primarily on the fact that the subject of this Congress is mathematics, which has been respected as "the queen of sciences" for its brilliant intellectual accomplishments, as suggested by the examples of the discovery of Goedel's theorem and the proof of the Fermat Last theorem in the last century. Mathematics is also "the servant of sciences" as explained by the great German mathematician Gauss when he spoke of "the queen of sciences". In the past century the application of mathematics witnessed rapid and more exciting development. The highly abstract languages, structures, methods and ideas created by mathematicians have been repeatedly proven to be universal instruments useful to other fields of science and technology and to economic and social development. This truly reflects the marvelous and close relations between mathematical theories and the objective world. Just by mentioning Riemann geometry and the theory of Relativity, Turing machine and the real computers, Radon integral and the CT scanners, we can see that mathematics is exerting more and more important influence on the modern civilization and social progress.

China had created glorious scientific and technological achievements in ancient times before a decline set in some three or four centuries ago. In 1915, the first Chinese comprehensive scientific society — "the Chinese Society for Science" was founded. Its founders were a group of students studying abroad, including a mathematician who was the first Chinese Ph.D. in mathematics. Starting with only 180 members at the beginning, the seeds it sowed are blossoming and bearing fruits in China today. The reform and opening up policy that China has adopted since 1978 has given tremendous impetus to the country's science and education. We have built up a well distributed system of research and a network of academic societies. Our scientists are working on many frontier projects in various fields. In the past 20 years Chinese scientists succeeded in constructing the electron-positron

collider, developing large computers and strong laser light sources, breeding hybrid rice and determining genetic codes, developing sophisticated word processing systems for Chinese characters, and setting up terrestrial stations for satellite remote sensing and nation-wide network for ecology observations. In mathematics, Chinese scholars have achieved important results in fields such as number theory, theorem-proving by computer, differential geometry, topology, complex analysis, probability and mathematical statistics, PDE, functional analysis, numerical analysis and control theory and so on.

Today, we have entered a new age, in which the social development is more dependent than ever before on the advancement of knowledge. This situation has brought about both opportunities and challenges to the development of science and technology in China. We have to work hard to keep pace with science and technology development in the world and strive to make greater contributions to the progress of human society. Science is an international endeavor, and no nation could be successful in isolation. International exchanges and cooperation in mathematics is of greater significance. As a universal language of science, mathematics plays a unique role in merging diverse cultures on the Earth. A typical example is the transmission of the oriental decimal numeration and the Greek geometry in history. I hope sincerely that the first International Congress of Mathematicians in the 21st century will open a new page in the history of world cultural exchanges. We will continue our efforts to promote international cooperation in science and technology.

In conclusion, I wish the Congress a great success, and hope that you all enjoy your stay in Beijing.

Thank you very much.

At recess — delegates are finding their own countries' locations on a stone terrestrial globe

Wu Wen-Tsun

President of the ICM2002

Ladies and Gentleman,

Sixteen years ago I attended as an observer on behalf of the Chinese Mathematical Society the 10th General Assembly of the International Mathematical Union in Oakland, at which CMS became a member of the IMU. I am very happy to see that the cooperation between Chinese mathematicians and the international mathematical community has been developing rapidly and fruitfully since then, and the inspiring progress is demonstrated today by the opening of the 24th ICM in Beijing. It is a high privilege and an honor for me to extend to you my warmest welcome.

Our science-mathematics, is an age-old yet evergreen field of human knowledge. The vitality of mathematics is, it seems to me, from its dealing with the numerical relation and spatial form in the most general sense. Numbers and forms, in the final analysis, reflect the most essential characters of things in the actual world. It is therefore no strange that the abstract theories and methods investigated by mathematicians would pervade almost all fields of science and technology. "Each science", as pointed out by Karl Marx, "could be considered to be perfect only if it permits the successful application of mathematics".

Mathematics gives, directly or indirectly, impetus to the development of productive forces as well. I mention here only one example — the revolutions of the communication industry, which would not have been possible without the mathematical physics from Gauss to Maxwell, and more recently without Turing and von Neumann's ideas of computers. It is therefore not without reasons that Napoleon has once said "the advancement and perfection of mathematics are intimately connected with the prosperity of the State". I prefer to quote again non-mathematician's viewpoint on the value of mathematics to avoid arousing suspicion of mathematicians' boast.

We are at the beginning of a new century. The unique situation of mathematics, different from any previous century at the turn, appears to be caused by the impact of the computers. Computers provide new tools, raise new problems, and allow new applications of mathematics. All that, I believe by my own research experience, will make a genuine new century of mathematics. It might be more challenging and promising to Chinese mathematicians whose country is struggling for transition from a developing society to the information and knowledge-based society.

Modern mathematics has historical roots of diverse civilizations. Mathematical activities in ancient China can be traced back to early time. The major pursuit of the ancient Chinese mathematicians was to solve problems expressed in equation. Along this line they contributed the decimal place-value numeration, negative and irrational numbers, various techniques for solving equations?etc. It is believable that ancient Chinese mathematicians had active knowledge exchanges with middle Asia and even Europe through the Silk Road. Today we have railways, airlines and even information highway instead of the Silk Road, the spirit of Silk Road-knowledge exchanges and cultural mergence ought to be greatly carried forward. I hope that the International Congress of Mathematicians 2002, held for the first time in a developing country, will open a glorious new page in the universal cooperation of mankind and bring with a prosperous future of our mathematical sciences.

I wish the Congress a success, wish you all a nice stay in Beijing.

Entertainment — Peking Opera performance

Liu Qi

Mayor of Beijing

Dear Delegates, Dear Guests, Ladies and Gentlemen:

Good afternoon! Today, I feel very honored to be present at International Congress of Mathematicians 2002. Here, on behalf of Beijing Municipal Government and the thirteen million people of the city, I would like to extend my sincere congratulations to the opening of this congress and express my warm welcome to scientists and guests participating in the conference.

ICM is committed to the research in one of the most basic disciplines of human knowledge. The intellectual fruits achieved in the field by mathematicians exert far-reaching influences on the progress of science and technology of human society and on the development of social culture and people's way of life. The fact that this conference is the first of its kind in the new century and the first session ever held in a developing country has given special significance to this meeting. The Municipal Government and myself are very pleased to be able to provide support and service to the meeting and we wish to present our highest compliments to mathematician and their exploration of reason.

The mathematic tradition in Beijing can be traced back to ancient times. Since the end of the nineteenth century, Beijing has played an important role in promoting the scientific and cultural exchanges between the east and the west. The city has nurtured numerous brilliant mathematicians, from Zhu Shijie in the thirteenth century to professor Chen Xingshen who is present here today. Now, Beijing continues to maintain its position as China's major center of mathematic education and research. Some two thousand mathematicians from the mathematic departments of tens of universities and research institutions such as Chinese Academy of Sciences are engaged in the education and high-level research of the field in an all-round way. At the same time, they keep extensive and close contacts and cooperate with their colleagues from countries and regions around the world.

Isn't it a pleasure to have friends from afar! The ancient and modern city continues its three thousand years history of civilization and composes its ode to the 2008 Olympics. We sincerely welcome you to tour around the city during your spare time. The city's historical monuments and sites will demonstrate you the charm of Chinese culture. The rapid development will bring your thoughts to the future of an international metropolis. I hope that all the guests will have a pleasant and efficient stay in Beijing and a beautiful memory in your heart.

May the conference a complete success! Thank you.

Ma Zhiming

Chairman of the Organizing Committee of ICM2002
President of the Chinese Mathematical Society

Ladies and Gentlemen,

After four years of preparation, the 24th International Congress of Mathematicians is now opening. It is my great honor on behalf of the Local Organizing Committee and the Chinese Mathematical Society to welcome you all to the ICM2002 in Beijing.

Four years is long for expecting, but short for preparing. Since the 13th General Assembly of the International Mathematical Union in Dresden in 1998, at which Beijing was chosen as the site of ICM2002, Chinese mathematical community has been racing against time to work for today's ICM2002. The first step was the setting up of the Local Organizing Committee in September of mathematical 1998, right after the Berlin Congress. The Committee, consisting of representatives from Taiwan, Hong Kong and overseas Chinese mathematicians, has been cooperating closely with the Executive Committee of IMU to ensure a smooth and effective preparation of this Congress. The preparation of the Congress is a symphony of international cooperation. I would like to take this opportunity to thank colleagues world-wide who have rendered all kinds of help and assistance. I am indebted in particular to IMU President Jacob Palis, Past President David Mumford, and Secretary Phillip Griffiths for their all-out support. Special thanks goes also to my German predecessor Professor Martin Groetschel, whose experiences of organizing the Berlin Congress are really helpful to us. The preparation of the Congress has won wide social and governmental support in China. The support from the government is evidenced by the presence of President Jiang Zemin and other Chinese leaders at this opening ceremony. The financial support from the Chinese government was even more than expected. The Organizing Committee of ICM2002 is grateful to the Chinese ministries and agencies that were listed on the slide shown left, the total of their funding is 10 million Chinese yuan, which amounts to about 1.2 million US dollars.

The spirit for the ICM2002 has been high among the Chinese public. Many Chinese scholars, teachers, industrialists, and even students were eager to contribute not only to help to prepare a successful ICM, but also to make the Congress a new start point for development of mathematics in China. Regarding the donations only, the Organizing Committee has received contributions of 3 million Chinese yuan from universities, industries and individuals. This amount is significant in

view of that China is still a developing country. Please watch the slides at left, which show the major donors, and I, in the name of the Organizing Committee of ICM2002, would like to extend to them our sincere thanks.

While the financial support is important, the scientific program is always the core of the Congress. Thanks to the Program Committee headed by Professor Y. Manin and the 19 international panels, the selected 20 plenary lectures and 174 invited lectures will, I believe, represent the latest advancement and frontier achievements in our science. The lectures given by the newly awarded Fields medallists and winner of the Nevanlinna Prize will of course highlight the scientific program of the Congress. On the other hand, more than 1200 short communications and poster presentations arranged by the local scientific committee will reflect the widespread active participation in the development of mathematics in recent years.

Up to now, the ICM2002 has 4,270 registered participants from 101 countries and regions, among whom 1 percent are from Australia, 3 percent from Africa, 56 percent from Asia, 16 percent from America, 24 percent from Europe. As the ICM held for the first time in a developing country, we see from above statistics that the percentage of the participation of mathematicians from developing countries is above 52 percent. The success of the financial program enabled us to make good our promise by various means to support financially about 400 scholars from developing countries and Eastern Europe (here I should thank the IMU for covering international traveling expenses for approximate 200 participants who are young mathematicians from developing countries and mathematicians from Eastern Europe, Africa and Latin America). In addition, the Organizing Committee has supported a number of mathematicians from western part of China as well.

Keeping in mind that it is the first ICM of the 21st century, the Organizing Committee has paid due attention to the programs for the general public, and considered it to be important for a new information era to attract the public to modern mathematics. Public talks on a range of topics and special activities related to the Congress were arranged for that purpose. Part of them are shown on the slide, among which I would like to mention here two examples: the Juvenile Mathematics Forum and the ICM2002 Mathematics Summer Campus, both were organized to raise the enthusiasm of young generation to mathematics that may have impact on the future of mathematics.

The 46 satellite conferences form a landscape of ICM2002. The slides show the list of satellite conferences, which are distributed geographically over 26 cities in different parts of China as well as 6 cities in Japan, Russia, Singapore, South Korea and Viet Nam. Almost for each satellite conference there is a story of international cooperation, the participation in of a number of Fields medallists, winners of Wolf Prize and winners of Nobel Prize made the whole program even more inspiring. Though it has been a tradition of ICMs to have a series of satellite conferences, the ICM2002 makes the satellite conference program broader in scale and more meaningful to a successful ICM. I would like therefore to express my thanks to all the local organizers of satellite conferences for their contribution.

Last but not the least, a few words about the logo of the ICM2002. The design was based on a diagram drawn by the 3rd century Chinese mathematician

Zhao Shuang to demonstrate Pythagoras theorem that appeared in ancient China first in Zhou Dynasty (11th century B.C—3rd century B.C). Some inspirations were put in to transform it to our logo. Let me show quickly by the video how does it make sense. First, by opening the edge of the outer square and enlarging the square inside, it will symbolize that mind of mathematicians are open, and that China is open. Next, varying colors make the diagram more like a rotating pinwheel to symbolize the hospitality of Beijing people. (Pinwheel is a folk toy which you may see children in Beijing's hutong playing with and greeting you: "Welcome, welcome!") Welcome to ICM2002, welcome to Beijing. Let us join hands to lift the veil of a new epoch of mathematics. I wish the congress a great success, and wish you all pleasant stay in Beijing.

Reception

Presentation of the Fields Medals

President Jiang Zemin granted the 2002 Fields Medals to Laurent Lafforgue and Vladimir Voevodsky (from left to right: Vladimir Voevodsky, Jiang Zeming, Jacob Palis, Laurent Lafforgue)

Fields Medal

Presentation of the Rolf Nevanlinna Prize

Phillip A. Griffths granted the 2002 Rolf Nevanlinna Prize to Madhu Sudan (right)

Nevanlinna Medal

ICM 2002 · Vol. I · 29–37

Closing Ceremony

The closing ceremony of the Congress was held at the Beijing International Conference Center on Wednesday, August 28, 2002, staring at 4:30pm. Here are the speeches at the closing ceremony.

Jacob Palis

President of the International Mathematical Union

Dear Colleagues, Ladies and Gentlemen,

At this moment, we are closing one more International Congress of Mathematicians, the 24th of a series that started in 1897 in Zurich in a span of more than one hundred years.

Thus, it's time to try to respond to the questions: Is it worthwhile to have such a comprehensive Congress, covering an impressive array of areas of mathematics, with 20 plenary talks, 174 invited lectures and many short communications? Were the lectures well presented in trying to reach a large mathematical audience, avoiding technical details and in offering an overview of the themes discussed and the prospect for research in the future? Is the Congress still attractive to a significant number of mathematicians from all over the world? Has it been organized in a way that led to the presence of a magical atmosphere combining friendship and inspiration for creativity in mathematics?

We have posed so many difficult questions and yet we are absolutely certain that the answers are all very positive. Indeed, the echoes from the participants are overwhelming: The Congress was one of the best ever. The lectures provided, to a large extent, a grand vision of today's mathematics and its prospect for tomorrow.

About 4,300 colleagues from 101 countries were present, among whom 2,700 are foreigners. Jointly, IMU and the Local Organizing Committee have supported the participation of about 450 foreign mathematicians from developing countries. A substantial part of the IMU support came from its Special Development Fund, to which the following institutions have contributed in the period 1998-2002: American Mathematical Society, Mathematical Society of Japan, London Mathematical Society, Brazilian National Research Council, Société Mathématique de France and

Wiskundig Genootschap Netherlands. To them we express our best thanks. Therefore, it's time to look to the future with optimism and determination in the pursuit of our dreams, in search of beauty in mathematics and its use to well serve society.

It's time also to warmly thank the Local Organizing Committee for their wonderful job. I wish I could name all 300 volunteers engaged in the organization, but I have to content myself in citing only five: Pei Zhuan, Luo Yang, Bao Ying, Li Yingjie and Hong Weizhe. As a symbol of the fine administrative support, I want to mention Ms Guo Wei. Our highest appreciation goes to President Jiang Ze-min for honoring the Congress with his presence at the Opening Ceremony and for co-awarding the Fields Medals. Hopefully, such a gesture by the highest dignitary of the host country may become, from now on, a tradition in the ICMs. Also, to the Chinese Institutions for their remarkable support in so many ways. To the Program Committee, we offer our sincere gratitude for the superb work in their choice of speakers.

Now, I want to finalize my words by presenting the main results of the 14th General Assembly that took place in Shanghai and again remarkably well prepared.

The officers of the International Mathematical Union for 2003-2006 are as follows:

Executive Committee

President:	John M. Ball	United Kingdom
Vice-Presidents:	Jean-Michel Bismut	France
	Masaki Kashiwara	Japan
Secretary:	Phillip A. Griffiths	USA
Members at Large:	Andrey A. Bolibruch	Russia
	Martin Grötschel	Germany
	Zhiming Ma	China
	Ragni Piene	Norway
	Madabusi S. Raghunathan	India
Ex-officio:	Jacob Palis (Past President)	Brazil

Commission on Development and Exchanges (CDE)

Chair:	Paulo Domingos Cordaro	Brazil
Secretary:	C. Herbert Clemens	USA
Members at Large:	Hajer Bahouri	Tunisia
	Graciela L. Boente Boente	Argentina
	Shrikrishna G. Dani	India
	Gérard Gonzalez-Sprinberg	France
	Fazal M. Mahomed	South Africa
	Toshikazu Sunada	Japan
	Jiping Zhang	China

International Commission on Mathematical Instruction (ICMI)

President:	Hyman Bass	USA
Vice Presidents:	Jill Adler	South Africa
	Michèle Artigue	France
Secretary:	Bernard R. Hodgson	Canada
Members at Large:	Carmen Batanero	Spain
	Mary Elizabeth Falk de Losada	Colombia
	Nikolai Dolbilin	Russia
	Peter Lawrence Galbraith	Australia
	Petar Stoyanov Kenderov	Bulgaria
	Frederick K.S. Leung	Hong Kong

International Commission on the History of Mathematics (ICHM)

Members at Large:	Jeremy John Gray	United Kingdom
	Wenlin Li	China

The Executive Committee also designated the following members for its Committee on Electronic Information and Communications: Pierre Berard (France), Jonathan Borwein-Chair (Canada), John Ewing (United States), Martin Gröetschel-EC representative (Germany), Alejandro Joffre (Chile), Peter Michor (Austria), David Morrison (United States), and Alf van der Poorten (Australia).

Various resolutions were voted at the General Assembly. Particularly, I would like to mention four of them:

Resolution 1

The General Assembly resolves that the next meeting of the General Assembly will be held at a time and place conveniently linked to the International Congress of Mathematicians in Madrid, Spain, in 2006.

Resolution 2

The General Assembly expresses its gratitude to the Organizing Committee of ICM2002, chaired by Ma, Zhiming.

The General Assembly also expresses its gratitude to Li Ta-tsien for his hospitality reception and excellent arrangements at General Assembly meeting in Shanghai.

Resolution 4

The General Assembly gives especial thanks to Phillip Griffiths for his excellent work as Secretary to the IMU over the last four years assisted by Arlen Hastings and Linda Geraci. It also thanks the Institute for Advanced Study (IAS) for its generous support of the IMU secretariat over this period.

Resolution 7

Notwithstanding these times of heightened tension and security concerns, we urge a continuation of scientific exchange and publication. The IMU opposes efforts either by governments, organizations, or individuals to restrict contacts and interactions in the world mathematical community. Specifically, we oppose holding individual mathematicians liable for the actions of their governments. The IMU endorses the principles expressed in the Article 5 of the Statutes of the International

Council for Science - ICSU, as adopted at the 1998 General Assembly, that reads as follows: In pursing its objections in respect of the rights and responsibilities of scientists, ICSU, as an international non-governmental body, shall observe and actively uphold the principle of the universality of science. This principle entails freedom of association and expression, access to data and information, and freedom of communication and movement in connection with international scientific activities, without any discrimination on the basis of such factors as citizenship, religion, creed, political stance, ethnic origin, race, colour, language, age or sex. ICSU shall recognize and respect independence of the internal science policies of its National Scientific Members. ICSU shall not permit any of its activities to be disturbed by statements or actions of a political nature.

All the resolutions will be published in the IMU Bulletin.

Thank you very much.

Presidents of the IMU: J. Ball and J. Palis

John Ball

President of the IMU for 2003-2006

Ladies and Gentlemen, Colleagues and Friends,

It is a great privilege to be elected as the next President of IMU and thus to have the opportunity, with the new Executive Committee, of helping to influence some of the important developments that will affect the mathematical community over the next few years.

It is a particular honour to succeed Jacob Palis, who for the last 12 years has held high office in IMU, for 8 years as Secretary and since 1999 as President. All those who know him will testify to the great energy, dedication and love for the community that he has brought to these posts.

This has been a memorable Congress, and we all recognize the very large number of people whose work has contributed to its outstanding success, those who served on the various international committees, the speakers for the many inspiring lectures, and above all the local organizers from Ma Zhiming through to the splendid student volunteers.

However, I would like to reserve some special words for the President of the Congress Professor Chern Shiing Shen. Despite his great age he was instrumental in ensuring the strong backing of the Chinese government for the Congress, and in his speech at the Opening Ceremony, and at other occasions during the Congress, he demonstrated the wisdom, warmth and dignity which are his hallmark. Professor Chern had hoped to attend the Closing Ceremony, but could not do so. But I am sure that his colleagues will convey to him our appreciation, not only for his contributions to this Congress, but also for his remarkable influence on our subject.

In addition to its traditional tasks, the new Executive Committee has much work to do. First there are important issues identified and developed through the work of the previous Executive Committee and IMU Committees, such as the project to retro-digitize the entire mathematics literature. Second, the General Assembly in Shanghai gave strong encouragement to the new Executive Committee to examine all the procedures and activities of the Union, and to report back to National Committees. And if I can mention one area to which I am personally committed, it is to see how IMU can better serve the needs of poorer and developing countries.

I can promise you that we will work hard, and with the help of mathematicians everywhere I hope that we will have some progress to report on when we meet again in Spain in 2006.

Thank-you.

Carles Casacuberta

Delegate of the Spanish IMU Committee

Dear Colleagues, Ladies and Gentlemen,

On behalf of the Spanish IMU Committee, which represents the Spanish mathematical societies, I am very pleased to invite you all to the next ICM, to be held in Madrid in 2006. The General Assembly of the IMU will be held just before, in Santiago de Compostela. Madrid is the capital of Spain, and Santiago, a UNESCO world heritage site, is in Galicia, in the northwest corner of the country.

We are well aware of the amount of work that the preparation of these events involves. We have been deeply impressed by the commitment of the organizers of the ICM2002 in Beijing, and also by the high level of the lectures and presentations. We hope to maintain these standards and are lucky to have the backing of the Spanish mathematical community and the help of many institutions. His Majesty the King of Spain has expressed his support for the venture in a letter that we received last week. Government and local authorities are also firmly behind us as we take on the responsibility for organizing the next ICM.

In addition, we plan to seek the cooperation of mathematical societies in Latin America. The event will provide an ideal opportunity to strengthen the links between countries and to create new channels for exchanges and joint work.

We are very grateful to the organizers of this ICM for allowing us to address this closing ceremony. Let me say warmheartedly: (*Hasta la vista*).

L. Faddeev

St. Petersburg Department of Steklov Mathematical Institute
Russian Academy of Sciences, Russia

Dear colleagues,

I think, that I can consider myself as a veteran of the International Mathematical Congresses. Indeed, the first one which I attended was the ICM1962 in Stockholm. At that time Soviet Academy decided to include several young researchers in the delegation and I was among them. Afterwards I was present on almost all ICM-s with exception of Vacouver 1974 and Helsinki 1978. My main impression here is that Beijin 2002 is one of the best Congresses both scientifically and organisationally.

The main idea of the ICM is to confirm the unity and universality of Mathematics. This Congress gave a lot of examples of this. Take for instance the sections of logic, number theory and algebra. The general underlining mathematical structures as well as language, used by speakers, were essentially identical.

I am highly impressed by the support to mathematics and fundamental science in genersl here in China. This is a great envy for many of us, coming from countries, where science and its needs are neglected. It was nice to realise that Russian mathematical school was highly represented here in Beijin. I would like to express my deep gratitude to organisers for the generous help, which allowed the presence at the Congress of many participants, living now in Russia and other countries of the FSU.

Thank you.

Abderrahman Boukricha

Faculte des Sciences de Tunis

Speech on behalf of the grantees of the Special Development Funds

The President of the International Mathematical Union,
The Chairman of the Local Organizing Committee,
Dear Colleagues, Ladies and Gentleman,

On behalf of financial grantees to the ICM2002, I would like to express our gratitude to the International Mathematical Union (IMU) for the travel support as well as The Local Organizing Committee for local expenses support.

We have really enjoyed our stay in Beijing and we are particularly grateful for all the exposure to the most recent development in various area of the mathematical sciences which reinforce the right way to the universal language and the universal knowledge.

We are also happy about the pleasant atmosphere as well as the friendliness and hospitality of the Chinese people.

Being here has also afforded me the opportunity of informing members of the ICM congress about the forthcoming Pan African Congress of Mathematicians scheduled to take place in Tunisia in September 2004.

Thank you very much.

Outside Door Party

Ma Zhiming

Chairman of the Organizing Committee of the ICM2002
President of the Chinese Mathematical Society

Ladies and Gentlemen, Dear colleagues:

You may remember the last words in my speech at the Opening Ceremony:

"Let us join hands to lift the veil of a new epoch of Mathematics. I wish the Congress a great success, and wish you all pleasant stay in Beijing."

At this moment I am very happy to say that what we expected has been achieved. As pointed out by the previous speakers at this Closing Ceremony, we have had a great success of the International Congress of Mathematicians 2002. I would therefore like to take this opportunity to thank all the institutions, organizations and individuals who have made efforts and contribution to ensure the success of the Congress.

First of all, I am grateful to all our participants coming from all over the world, your enthusiastic participation offered a major guarantee of the success of the Congress. Let me express once again, as I did at the Opening Ceremony, our gratitude to the broad social organizations and governmental ministries and to IMU for their valuable support, without such support there would have been no success of the Congress. Special thanks go also to all our invited speakers for their remarkable lectures which represented the latest advancement and frontier achievements in our science and marked really a high academic level of our Congress. The three public lectures attracted a broad social audience and were of great significance to the popularization of mathematics and its applications. Also I would like to mention that the short communications and poster presentations arranged by the local scientific committee reflected the wide and active development of mathematics in recent years.

I have a long list of Chinese organizations and colleagues whom we should appreciate for their contribution towards the success of the Congress. Because of the time limitation I could not mention all their names here, but we shall never forget their excellent work.

Let me conclude my speech with sincere thanks to you all again and with best wishes for a new golden age of our science of mathematics.

I declare the 24th International Congress of Mathematicians closed.

ICM 2002 · Vol. I · 39–85

List of Participants

Abate, Marco (Italy)
Abd Al-Kader, Gamal (Egypt)
Abdollahi, Alireza (Iran)
Abdounur, Oscar Joao (Brazil)
Abel, Mart (Estonia)
Abel, Mati (Estonia)
Abeles, Francine F. (USA)
Abrahamsson, Leif (Sweden)
Abreu-Blaya, Ricardo (Cuba)
Adamczewski, Boris (France)
Adem, Alejandro (USA)
Adhikari, S. D. (India)
Afanasyev, Alexander Petrovitch (Russia)
Afshar Nejad, Zahra (Iran)
Aftalion, Amandine (France)
Aghdasi Alamdari, Seyedeissa (Iran)
Agrachev, Andrei A. (Italy)
Ahn, Soon Jeong (Korea)
Ahn, Youngho (Korea)
Ahuja, Mangho (USA)
Ai, Mingyao (China)
Ai, Sumei (China)
Ai, Wenbao (China)
Aiba, Akira (Japan)
Aikawa, Hiroaki (Japan)
Aitchison, Iain R. (Australia)
Akbari, Saieed (Iran)
Akbary, Amir (Canada)
Akhmetèv, Peter Mikhailovich (Russia)
Akhmetov, Denis Robertovich (Russia)
Akira, Mizutanz (Japan)
Aksoy, Asuman G. (USA)
Alaeiyan, Mehdi (Iran)
Alamatsaz, Mohammad H. (Iran)
Alarcón, Francisco E. (USA)
Albeverio, Sergio (Germany)
Albouy, Alain (France)
Alefeld, Goetz E. (Germany)
Aleksandrov, Aleksandr Grigorjevich (Russia)
Alesker, Semyon (Israel)
Alestalo, Pekka (Finland)
Alif, Mohssine (Morocco)
Alldridge, Alexander (Germany)
Allman, Elizabeth S. (USA)
Almocera, Lorna S. (Philippines)
Alon, Noga (Israel)
Alonso, Ana I. (Spain)
Alrashed, Maryam H. A. (United Kingdom)
Alshehri, Mohammed (Saudi Arabia)
Alshin, Alexander (Russia)
Alshina, Elena (Russia)
Altay, Sezgin (Turkey)
Altman, Allen (USA)
Altshuler, Amos (Israel)
Alves, Manuel J. (Mozambique)
Ambrosio, Luigi (Italy)
Amini, Siamak (United Kingdom)
Amirali, Gabil M. (Turkey)
Amiraliyeva, Ilhame (Turkey)
Amitani, Yasuharu (Japan)
Amornsamankul, Somkid (Thailand)
Amyari, Maryam (Iran)
An, Fengwen (China)
An, Guimei (China)
An, Hengpin (China)
An, Hongzhi (China)
An, Jinpeng (China)
An, Tianqing (China)
An, Xinming (China)
Anand, Nukala (India)
Anbu Durai, M. (India)
Ancel, Fredric D. (USA)
Ancona, Vincenzo (Italy)

Ando, Shiro (Japan)
Andradas, Carlos (Spain)
Andreev, Vsevolod V. (Russia)
Andrews, Benjamin H. (Australia)
Andrey, Ladislav (Czech)
Anichini, Giuseppe (Italy)
Aoki, Shigeru (Japan)
Apiwattanapong, Supasit (Thailand)
Appell, Jurgen (Germany)
Araki, Huzihiro (Japan)
Arceo, Carlene P. (Philippines)
Archibald, W. Thomas (Canada)
Argerami, Martin (Argentina)
Aripov, Mersaid (Uzbekistan)
Armas-Sanabria, Lorena (Mexico)
Arnold, Douglas N. (USA)
Arnold, Vladimir (Russia)
Arora, Sanjeev (USA)
Artamonov, Nikita Vyacheslavovich (Russia)
Artamonov, Viatchesalv (Russia)
Arvet, Pedas (Estonia)
Arzhantseva, Goulnara N. (Switzerland)
Asada, Teruko (Japan)
Asano, Hiroshi (Japan)
Asanov, A. (Kyrgyzstan)
Asavanant, Jack (Thailand)
Asawakun, Prapasri (Thailand)
Asgari, Mahdi (USA)
Ash, J. Marshall (USA)
Ashino, Ryuichi (Japan)
Ashiq, Muhammad (Pakistan)
Ashna, Amir Hossein (Iran)
Astley, Roger S. (USA)
Aulbach, Bernd (Germany)
Auroux, Denis (France)
Avdispahic, Muharem (Bosnia and Herzegovina)
Avgerinos, Evgenios P. (Greece)
Avramidi, Ivan G. (USA)
Avritzer, Dan (USA)
Ayaragarnchanakul, Jantana (Thailand)
Ayoub, Ayoub B. (USA)
Ayupov, Abdullaevich (Uzbekistan)
Ayupov, Shavkat A. (Uzbekistan)
Azarpanah, Fariborz (Iran)
Azimi, Parviz (Iran)
Bacopoulos, Alex (Greece)
Badiozzaman, Abdul Jabbar (Iran)
Bae, Wonsoung (Korea)
Bagchi, Atish (USA)
Bahouri, Hajer (Tunesia)
Bai, Chengming (China)
Bai, Fengshan (China)
Bai, Jie (China)
Bai, Ruipu (China)
Bai, Shizhong (China)
Bai, Z.J. (USA)
Bai, Zhongzhi (China)
Baizhanov, Bektur Sembiuly (Kazakhstan)
Bak, Anthony (Germany)
Baklouti, Ali (Tunesia)
Balashova, Galina S. (Russia)
Ball, Deborah Loewenberg (USA)
Ball, John M. (United Kingdom)
Balmaceda, Jose Maria P. (Philippines)
Banaszak, Grzegorz Marian (Poland)
Banchoff, Thomas F. (USA)
Banihashemi, Saeed Seyed Agha (Iran)
Banulescu, C. Martha (Romania)
Bao, Jiguang (China)
Bao, Weihua (China)
Bao, Zhiqiang (China)
Baoulina, Ioulia (Russia)
Barany, Imre (Hungary)
Barberis, Maria Laura (Argentina)
Barrett, David E. (USA)
Barrientos, Aniura Milanes (Brazil)
Barron, Katrina D. (USA)
Bartnik, Robert (Australia)
Bartolone, Claudio G. (Italy)
Bashkirtseva, Irina (Russia)
Bass, Hyman (USA)
Bau, Sheng (South Africa)
Bav, Sheng (South Africa)
Bayasgalan, Tsembeltsogt (Mongolia)
Beattie, Margaret A. (Canada)
Beattie, Ronald J. (Canada)
Beesham, Aroonkumar (South Africa)

Behforooz, Hossein (USA)
Behrend, Kai (Canada)
Behrends, Ehrhard (Germany)
Behzad, Mehdi (Iran)
Belinsky, Eduard (Barbados)
Belousov, Evgeny (Russia)
Belov, Ilia (Russia)
Ben Arous, Gérard (Switzerland)
Ben Taher, Rajae (Morocco)
Benedicks, Michael (Sweden)
Beniash-Kryvets, Valery V. (Belarus)
Berg, Christian (Denmark)
Bergeron, Nicolas (France)
Berkovich, Lev M. (Russia)
Berkovits, Juha T. (Finland)
Bertoin, Jean R. (France)
Bestvina, Mladen (USA)
Bhadra Man, Tuladhar (Nepal)
Bhandari, A.K. (India)
Bhattacharya, Siddhartha (India)
Bhattacharyya, P. (India)
Bhattacharyya, Rabindra Kumar (India)
Bhattacharyya, Sandip (India)
Bhattarai, Hom Nath (Nepal)
Bhatwadekar, S. M. (India)
Bi, Jianxing (China)
Bi, Sikun (China)
Bian, Baojun (China)
Biane, Philippe (France)
Bickel, Peter J. (USA)
Biegler-Koenig, Friedrich (Germany)
Bigelow, Stephen J. (Australia)
Bimonte, Giuseppe R. (Italy)
Bin, Honghua (China)
Biran, Paul Ian (Israel)
Birnir, Bjorn (USA)
Biscari, Paolo (Italy)
Bisch, Dietmar Herbert (USA)
Bismut, Jean-Michel (France)
Biss, Daniel K. (USA)
Blackadar, Bruce E. (USA)
Blanco, Maria F. (Spain)
Bland, John S. (Canada)
Bleiler, Steven A. (USA)
Blizorukov, Michael G. (Russia)
Bloemker, Dirk (Germany)
Blokhuis, Aart (Netherlands)
Bo, Ling (China)
Bodnarescu, Musat V. (Germany)
Boersema, Jeff (USA)
Boffi, Daniele (Italy)
Bognár, Gabriella (Hungary)
Bolibrukh, Andrey (Russia)
Boling, Patricia M. (USA)
Bollobas, Bela S. (USA)
Bolt, Michael D. (USA)
Bolthausen, Erwin (Switzerland)
Bonatti, Christian (France)
Bondal, Alexei I. (Russia)
Bonk, Mario (USA)
Booss-Bavnbek, Bernhelm (Denmark)
Borisenko, Alexander (Ukraine)
Borisov, Denis (Russia)
Borovoi, Mikhail (Israel)
Borwein, Jonathan M. (Canada)
Borzooei, Rajab Ali (Iran)
Bos, Henk J.M. (Netherlands)
Bottazzini, Umberto (Italy)
Bouarroudj, Sofiane (Japan)
Bouchiba, Samir (Morocco)
Boukricha, Abderrahman (Tunesia)
Boulton, Lyonell (Venezuela)
Bourguignon, Jean-Pierre (France)
Bourhim, Abdellatif (Morocco)
Bouscaren, Elisabeth (France)
Bouzar, Chikh (Algeria)
Brännström, Äke N. (Sweden)
Brambila-Paz, Leticia (Mexico)
Brandao, Daniel Smania (USA)
Branner, Bodil (Denmark)
Bratkov, Yuri N. (Russia)
Braun-Angott, Peter (Germany)
Bray, Hubert L. (USA)
Bredimas, Antoine-A.P. (Greece)
Brenier, Yann (France)
Brenner, Sheila (United Kingdom)
Bressan, Alberto (Italy)
Bricmont, Jean L. (Belgium)
Brieden, Andreas (Germany)
Bronstein, Manuel (France)

Brown, Lawrence D. (USA)
Brown, Robert F. (USA)
Bruce, Gilligan (Canada)
Brunner, Hermann (Canada)
Brychkov, Yury (Canada)
Brzezniak, Zdzislaw (United Kingdom)
Bshouty, Daoud (Israel)
Bu, Charles Qiyue (USA)
Bu, Shangquan (China)
Bufetov, Alexander I. (USA)
Bujalance, Emilio (Spain)
Bukhari, Abdul-Fattah (Saudi Arabia)
Buldaev, Alexander (Russia)
Bulinski, Alexander V. (Russia)
Bullett, Shaun (United Kingdom)
Burger, Reinhard (Austria)
Bursztyn, Henrique (Belgium)
Buskes, Gerard J. (USA)
Butkovic, Davor (Croatia)
Butler, Michael C.R. (United Kingdom)
Byrne, Catriona M. (Germany)
Caceres, Luis (USA)
Caffarelli, Luis Angel (USA)
Cai, Jichuan (China)
Cai, Jinxing (China)
Cai, Kai (China)
Cai, Kairen (China)
Cai, Maocheng (China)
Cai, Yingchun (China)
Caicedo, Xavier (Colombia)
Calderer, Maria-Carme T. (USA)
Camacho, Luisa M. (Spain)
Cangul, Ismail Naci (Turkey)
Canino, Annamaria (Italy)
Cannon, John J. (Australia)
Cao, Chongguang (China)
Cao, Chunlei (China)
Cao, Daomin (China)
Cao, Feilong (China)
Cao, Guangfu (China)
Cao, Hongjun (China)
Cao, Hui (China)
Cao, Jiading (China)
Cao, Jianwen (China)
Cao, Lei (China)
Cao, Linfen (China)
Cao, Liqun (China)
Cao, Rui (China)
Cao, Wei (China)
Cao, Xian-Bing (China)
Cao, Xiaohong (China)
Cao, Xiwang (China)
Cao, Yan (USA)
Cao, Yonglin (China)
Cao, Zhenfu (China)
Cardoso, Fernando F. (Brazil)
Carrascal, Alexander S. (Philippines)
Casacuberta, Carles (Spain)
Casadio Tarabusi, Enrico (Italy)
Casas, Eduardo (Spain)
Cascaval, Radu C. (USA)
Cassaigne, Julien (France)
Cassy, Bhangy (Mozambique)
Castellanos, Victor (Mexico)
Catepillan Hearn, Ximena (USA)
Cawagas, Raoul Enrico (Philippines)
Cayford, Afton H. (Canada)
Cen, Xuejuan (China)
Cerami, Giovanna M. (Italy)
Chahal, Jasbir S. (USA)
Chai, Young Do (Korea)
Chalub, Fabio Augusto (Austria)
Chan, Tony F.C. (USA)
Chang, Chiu Cheng (China)
Chang, Gengzhe (China)
Chang, In Kap (Korea)
Chang, Jen (Canada)
Chang, Keqi (China)
Chang, Kun Soo (Korea)
Chang, Kung-Ching (China)
Chang, Qianshun (China)
Chang, Sun-Yung Alice (USA)
Chang, Wenwu (China)
Chang, Xiao (China)
Chang, Yanxun (China)
Chang, Yaotsu (Taiwan)
Chang, Yu (China)
Chao, Xiaoli (China)
Chasuriya, Pachara (Thailand)
Chatterji, Srishti D. (Switzerland)

Chaudhry, Muhammad Aslam (Saudi Arabia)
Chaudhuri, Kripasindhu (India)
Chekanov, Yuri V. (Russia)
Chemia, Karine (France)
Chen, Anyue (United Kingdom)
Chen, Beifang (China)
Chen, Bo (China)
Chen, Boyong (China)
Chen, Caisheng (China)
Chen, Chuanmiao (China)
Chen, Chuanping (China)
Chen, Chuanzhong (China)
Chen, Dayue (China)
Chen, Dechang (USA)
Chen, Erming (China)
Chen, Fangqi (China)
Chen, Fengde (China)
Chen, Gu (China)
Chen, Guangyu (China)
Chen, Gui-Qiang (USA)
Chen, Guoliang (China)
Chen, Guolong (China)
Chen, Hanfu (China)
Chen, Hanlin (China)
Chen, Hao (China)
Chen, Hua (China)
Chen, Huangen (China)
Chen, Jiading (China)
Chen, Jiahong (China)
Chen, Jianhua (Beijing)(China)
Chen, Jianhua (China)
Chen, Jianlong (China)
Chen, Jianqing (China)
Chen, Jiecheng (China)
Chen, Jinglin (China)
Chen, Jinru (China)
Chen, Jinwen (China)
Chen, Joseph Cheng-Yih (USA)
Chen, Junjie (China)
Chen, Kuo-Chang (USA)
Chen, Lansun (China)
Chen, Li (China)
Chen, Liangyun (China)
Chen, Lianzhi (China)
Chen, Linda (USA)
Chen, Louis H.Y. (Singapore)
Chen, Lusheng (China)
Chen, Meng (China)
Chen, Min (China)
Chen, Mingyu (China)
Chen, Mu-Fa (China)
Chen, Peide (China)
Chen, Peifang (China)
Chen, Qingtao (China)
Chen, Rongsi (China)
Chen, Shaohua (Canada)
Chen, Sheng (China)
Chen, Shouquan (China)
Chen, Shunmin (China)
Chen, Shuping (China)
Chen, Shutao (China)
Chen, Shuxing (China)
Chen, Thomas (USA)
Chen, Tianping (China)
Chen, Wei (China)
Chen, Weihuan (China)
Chen, Wende (China)
Chen, Wenxiong (USA)
Chen, Wenyi (China)
Chen, William Y.C. (China)
Chen, Xia (USA)
Chen, Xiao (China)
Chen, Xiaoman (China)
Chen, Xiaosong (China)
Chen, Xin (China)
Chen, Xinkang (China)
Chen, Xinli (China)
Chen, Xiongchun (China)
Chen, Xiru (China)
Chen, Xiu Su (China)
Chen, Xiuhong (China)
Chen, Xiuxiong (USA)
Chen, Xuejuan (China)
Chen, Yemin (China)
Chen, Yifei (China)
Chen, Yihong (China)
Chen, Yin (Canada)
Chen, Yonggao (China)
Chen, Yong-Gao (China)

Chen, Yu (Italy)
Chen, Yun-Gang (Japan)
Chen, Yurong (China)
Chen, Yuyun (China)
Chen, Zeqian (China)
Chen, Zhangxin (USA)
Chen, Zhen-Qing (USA)
Chen, Zhi (Japan)
Chen, Zhiguo (China)
Chen, Zhihua (China)
Chen, Zhiming (China)
Chen, Zhiqi (China)
Chen, Zili (China)
Chenciner, Alain (France)
Cheng, Bin (China)
Cheng, Daizhan (China)
Cheng, Jian (China)
Cheng, Jin (China)
Cheng, Kuo-Shung (China)
Cheng, Lixin (China)
Cheng, Qing-Ming (Japan)
Cheng, Qiyuan (China)
Cheng, Rongjun (China)
Cheng, Wai-Yan (China)
Cheng, Wei (China)
Cheng, Xinyue (China)
Cheng, Yuanji (Sweden)
Cheng, Zhian (China)
Chern, I-Liang (Taiwan)
Chern, S. S. (China)
Chernyavskya, Nina (Israel)
Chertock, Alina (USA)
Chetverushkin, Boris (Russia)
Cheung, Yitwah (USA)
Chi, Dong Pyo (Korea)
Chien, Mao-Ting (Taiwan)
Chinchaladze, Natalia (Georgia)
Chinta, Gautam (USA)
Cho, Sanghun (Korea)
Cho, Sung Je (Korea)
Cho, Sung Mun (Korea)
Cho, Yong Seung (Korea)
Cho, Yunhi (Korea)
Choe, Boo R. (Korea)
Choe, Sun-Bong (DPR of Korea)
Choe, Young H. (Korea)
Choi, Jeongwhan (Korea)
Choi, Man-Duen (Canada)
Choi, Myungjun (Korea)
Choi, Q-Heung (Korea)
Choi, Yeontaek (Korea)
Chong, Yun Chol (DPR of Korea)
Chotchaisthit, Somchit (Thailand)
Choudhary, Bishweshwar (Zambia)
Choutiaev, Victor (Russia)
Chow, Shue S. (USA)
Chu, Eric K. (Australia)
Chu, Zhaofang (China)
Chumni, Wichai (Thailand)
Chundang, Ungsana (Thailand)
Chung, Dong Myung (Korea)
Chung, Jaeyoung (Korea)
Chung, Young-Bok (Korea)
Ciegis, Raimondas (Lithuania)
Cifventes, Particio (Spain)
Cioranesco, Doina M. (France)
Cipriani, Fabio E.G. (Italy)
Cipu, Mihai (Romania)
Cisneros-Molina, Jose L. (Mexico)
Clee, William A. (USA)
Clutterbuck, Julie (Australia)
Cogdell, James W. (USA)
Cohen, Albert (France)
Cohen, Arjeh M. (Netherlands)
Cohen, Henri J. (France)
Coleman, Robert Frederick (USA)
Collino, Alberto (Italy)
Combettes, Patrick Louis (France)
Cong, Fuzhong (China)
Constantinescu, Eliodor (Romania)
Constantinescu, Gabriela (Romania)
Conte, Alberto (Italy)
Contreras, Gonzalo (Mexico)
Cooper, Peter (United Kingdom)
Coornaert, Michel (France)
Cornuejols, Gerard (USA)
Corti, Alessio (United Kingdom)
Cristea, Valentin Gabriel (Romania)
Crnkovic, Dean (Croatia)
Croft, Hallard T. (United Kingdom)

Csirik, Janos A. (USA)
Csorgo, Piroska (Hungary)
Cui, Baotong (China)
Cui, Bin (China)
Cui, Chengri (China)
Cui, Guizhen (China)
Cui, Hengjian (China)
Cui, Jianlian (China)
Cui, Jinchuan (China)
Cui, Junzhi (China)
Cui, Lihong (China)
Cukrowicz, Jutta R. (Germany)
Cuntz, Joachim Jr (Germany)
Curran, Stephen (USA)
Da Luosang , Langjie (China)
Dadakhodjaev, R.A. (Uzbekistan)
Dafermos, Constantine M. (USA)
Dai, Binxiang (China)
Dai, Chaoshou (China)
Dai, Daoqing (China)
Dai, Haohui (China)
Dai, Jialing (China)
Dai, Qiang (China)
Dai, Qin (China)
Dai, Qingguo (China)
Dai, Wanyang (China)
Dai, Xianzhe (USA)
Dai, Xiongping (China)
Dajczer, Marcos (Brazil)
Dalawat, Chandan Singh (India)
Dale, Knut T. (Norway)
Damian, Florin (Moldova)
Dan, Avritzer (Brazil)
Dan, Nicusor (Romania)
Daneshkhah, Ashraf (Iran)
Dang, Trong Duc (Vietnam)
Dang, Yaru (China)
Danielyan, Arthur A. (USA)
Danila, Gentiana (France)
Dannan, Fozi Mustafa (Qatar)
Dara, Moazzami (Iran)
Darafsheh, Mohammad Reza (Iran)
Darus, Maslina (Malaysia)
Darwish, Mohamed Abdalla (Egypt)
Das, Manav (USA)
Dauben, Joseph W. (USA)
Daverman, Robert J. (USA)
Davidson, Kenneth R. (Canada)
De Blasi, Francesco Saverio (Italy)
De La Pena, Jose A. (Mexico)
De Loera Herrera, Jesus A. (USA)
De Medeiros, Nivaldo N. (USA)
De Queiroz, Marcelo Gomes (Brazil)
De Silva, Basil M. (Australia)
De Young, Gregg Ronald (Egypt)
Dean, Andrew J. (Canada)
Dean, Sarah E. (USA)
Debinska-Nagorska, Anna (Poland)
Debnath, Lokenath (USA)
Dehornoy, Patrick (France)
Del Rosario, Ricardo C.H. (Philippines)
Delorme, Patrick (France)
Demelo, Welington (Brazil)
Dementiev, Oleg N. (Russia)
Dementieva, Maria B. (Russia)
Demidenko, Gennadii V. (Russia)
Demmel, James Weldon (USA)
Dencker, Nils (Sweden)
Denef, Jan (Belgium)
Deng, Aijiao (China)
Deng, Bangming (China)
Deng, Guantie (China)
Deng, Jianping (China)
Deng, Mingli (China)
Deng, Mingxiang (China)
Deng, Pingji (China)
Deng, Yanjuan (China)
Deng, Yingchun (China)
Deng, Yingpu (China)
Deng, Yuping (China)
Deng, Zhimin (China)
Denisov, Alexander S. (Russia)
Derfel, Gregory A. (Israel)
Deshmukh, Kishor C. (India)
Desquith, Etienne (Ivory Coast)
Deszcz, Ryszard (Poland)
Deutsch, Jesse (Botswana)
Dev Sarma, Bijoy K. (India)
Devadoss, Satyan L. (USA)
Dewan, Kum Kum (India)

Dhamacharoen, Ampon (Thailand)
Dhungana, Bishnu P. (Korea)
Di, Jizheng (China)
Di, Yanjun (China)
Diacu, Florin N. (Canada)
Diagana, Youssouf Mahamadou (Ivory Coast)
Diansuy, Maria Ailynn A. (Philippines)
Dibaei, Mohammad Taghi (Iran)
Diblik, Josef (Czech)
Dickenstein, Alicia M. (Argentina)
Dierkes, Ulrich (Germany)
Dindos, Martin (USA)
Ding, Chunmei (China)
Ding, Fan (China)
Ding, Guanggui (China)
Ding, Jianmin (China)
Ding, Lijuan (China)
Ding, Qing (China)
Ding, Ren (China)
Ding, Shusen (USA)
Ding, Tongren (China)
Ding, Wanding (China)
Ding, Weiyue (China)
Ding, Xiaoping (China)
Ding, Xiaqi (China)
Ding, Xuanhao (China)
Ding, Yan (China)
Ding, Yanheng (China)
Ding, Yiming (China)
Ding, Yong (China)
Ding, Yuan Zhi (China)
Ding, Zhiguo (China)
Ding, Zhonghai (USA)
Djokovic, Dragomir Z. (Canada)
Djuraev, Tukhtamurad Djuraevich (Uzbekistan)
Dmitruk, Andrei V. (Russia)
Dobrev, Vladimir Krastev (Bulgaria)
Doi, Masaaki (Japan)
Donaldson, Simon (United Kingdom)
Dong, Jianze (China)
Dong, Junhua (China)
Dong, Qingwei (China)
Dong, Xiaolei (China)
Dong, Xinhan (China)
Dong, Xisong (China)
Dong, Xun (USA)
Dong, Yan (China)
Dong, Ye (France)
Dong, Yujun (China)
Dong, Yuxin (China)
Dong, Zhao (China)
Donoho, David L. (USA)
Doob, Michael (Canada)
Doplicher, Luisa (Italy)
Doplicher, Sergio (Italy)
Dorfmeister, Josef F. (Germany)
Dorfmeister, Josef G. (Germany)
Dorier, Jean-Luc (France)
Dostal, Zdenek (Czech)
Dotti, Isabel G. (Argentina)
Douglas, Michael Ronald (USA)
Douglas, Robert (USA)
Dovbinin, Vladimir (Russia)
Drabek, Pavel (Czech)
Dragovic, Vladimir (Yugoslavia)
Drozd, Yuriy (Germany)
Druck, Suely (Brazil)
Dryuma, Valerii (Moldova)
Du, Caifeng (China)
Du, Chaohe (China)
Du, Dongyun (China)
Du, Jinyuan (China)
Du, Nanlin (China)
Du, Qiang (USA)
Du, Qikui (China)
Du, Qing-Yan (China)
Du, Ruoxia (China)
Du, Shaofei (China)
Du, Yihong (Australia)
Du, Yiping (China)
Du, Yonghong (China)
Du Sautoy, Marcus P. F. (United Kingdom)
Duan, Haibao (China)
Duan, Jinqiao (USA)
Duan, San (China)
Duan, Wenying (China)
Duan, Xuefu (China)

Duan, Yongjiang (China)
Duan, Zeyong (China)
Dubovski, Pavel B. (Russia)
Duc Trong, Dang (Indonesia)
Duduchava, Roland (Georgia)
Duell, Wolf-Patrick (Germany)
Dumortier, Freddy (Belgium)
Dunajski, Maciej (United Kingdom)
Dung, Nguyen Viet (Vietnam)
Dunne, Edward G. (USA)
Duong, Duc Minh (Vietnam)
Duong, Vun. (France)
Duplij, Steven A. (Ukraine)
Durfee, Alan H. (USA)
Dusembaev, Anuar (Kazakhstan)
Dvornicich, Roberto (Italy)
Dzhenaliev, Muvasharkhan Tanabayevich (Kazakhstan)
E, Weinan (USA)
Ebrahimi-Vishki, Hamid Reza (Iran)
Eckmann, Jean-Pierre (Switzerland)
Edward, J.Robert Victor (India)
Edwards, Robert D. (USA)
Egorov, Alexandre A. (Russia)
Eida, Atsuhiko (Japan)
Eilers, Soren (Denmark)
Eisenbud, David (USA)
Ekhaguere, Godwin O. S. (Nigeria)
Ekong, Samuel (France)
El Hassan, El Kinani (Morocco)
El Karoui, Nicole (France)
El Sabaa, Fawzy Mohamed Fahmy (Egypt)
El-Afifi, Mohamed Mohamed Abdel Mageed (Egypt)
El-Bakry, Mahmoud Abdel-Rahman (Egypt)
El-Doma, Mohamed (Sudan)
Elena, Roubachkina (Russia)
Elettreby, Mohammed Fathy (Egypt)
Elkaroui, Nicole (France)
Elliott, George A. (Canada)
Elworthy, Kenneth David (United Kingdom)
Elyar, Gasanov (Russia)
Emerson, Annette W. (USA)
Emmer, Michele (Italy)
Engquist, Bjorn (Sweden)
Enkhbat, Rentsen (Mongolia)
Enock, Michel (France)
Enomoto, Kazuyuki (Japan)
Epple, Moritz (Germany)
Eremenko, Alexandre (USA)
Eriksson-Bique, Sirkka-Liisa (Finland)
Erjaee, G.Hossain (Iran)
Ershov, Yury L. (Russia)
Escaner, Jose Maria Iv L. (Philippines)
Esfandyari-Kalejahi, Abdolrasoul (Iran)
Eslahchi, Changiz (Iran)
Eslami, Ziba (Iran)
Esnault, Hélène (Germany)
Esslamzadeh, Gholam Hossein (Iran)
Etingof, Pavel I. (USA)
Eto, Kazufumi (Japan)
Eudave-Munoz, Mario (Mexico)
Evans, David E. (United Kingdom)
Evslin, Jarah M. (Italy)
Ewing, John H. (USA)
Faddeev, Ludwig Dmitrievich (Russia)
Fakharzadeh J., Alireza (Iran)
Falcone, Giovanni (Italy)
Fan, Ai Hua (China)
Fan, Bolin (China)
Fan, Dashan (USA)
Fan, Engui (China)
Fan, Genghua (China)
Fan, Huijun (China)
Fan, Lei (China)
Fan, Shaofeng (China)
Fan, Shuqin (China)
Fan, Suohai (China)
Fan, Tianyou (Beijing)(China)
Fan, Tian-You (China)
Fan, Yulian (China)
Fan, Zhoutian (China)
Fanaai, Hamidreza (Iran)
Fandom Noubiap, Roger (Germany)
Fang, Biqi (China)
Fang, Daoyuan (China)
Fang, Fuquan (China)

Fang, Hui (China)
Fang, Junsheng (China)
Fang, Qizhi (China)
Fang, Shao Mei (China)
Fang, Shizan (France)
Fang, Tongzhu (China)
Fang, Wang (China)
Fang, Xingui (China)
Fang, Yanbin (China)
Fang, Yong (China)
Fang, Youkang (Botswana)
Fang, Zhong Bo (Korea)
Faridi, Sara (USA)
Farr, Graham E. (Australia)
Fathi, Albert (France)
Favro, Ruth G. (USA)
Fedeli, Alessandro (Italy)
Fei, Li (China)
Fei, Ming Wen (China)
Fei, Yutian (China)
Feichtinger, Gustav (Austria)
Feige, Uriel (Israel)
Feireisl, Eduard (Czech)
Feng, Chun Rong (China)
Feng, Gang (China)
Feng, Huitao (China)
Feng, Jianfen (China)
Feng, Keqin (China)
Feng, Qi (China)
Feng, Qiuxiang (China)
Feng, Quandong (China)
Feng, Ruyong (China)
Feng, Shaoji (China)
Feng, Shui (Canada)
Feng, Tao (Sweden)
Feng, Wei (China)
Feng, Weijie (China)
Feng, Wenying (Canada)
Feng, Yongping (China)
Feng, Yuming (China)
Ferenczi, Sebastien (France)
Fernandez, Arturo (Spain)
Fernandez, Jose L. (Spain)
Ferreira, Fernanda Amelia (Portugal)
Ferreira, Fernando Flavio (Portugal)
Ferreira, Vitor O. (Brazil)
Ferrero, Daniela (USA)
Fiedler, Bernold (Germany)
Finashin, Sergey (Turkey)
Flajolet, Philippe (France)
Flatto, Leopold (USA)
Flicker, Yuval (USA)
Flores-Bazan, Fabian (Chile)
Flores-Espinoza, Ruben (Mexico)
Flytzanis, Elias (Greece)
Foellmer, Hans (Germany)
Fontaine, Jean-Marc (France)
Forni, Giovanni (USA)
Forti, Marco (Italy)
Forys, Iwona (Poland)
Fourati, Sonia (France)
Fourie, Jan H. (South Africa)
Freed, Daniel S. (USA)
French, Christopher P. (USA)
Friedman, Eduardo (Chile)
Frohardt, Daniel E. (USA)
Frolova, Julia (Russia)
Frommer, Henning (Germany)
Fu, Baohua (France)
Fu, Junyi (China)
Fu, Lei (China)
Fu, Siqi (USA)
Fu, Tung-Shan (Taiwan)
Fu, Xilin (China)
Fu, Yudong (China)
Fugère, Jean Benoit (Ecuador)
Fujiwara, Koji (Japan)
Fukai, Yasunari (Japan)
Fukuda, Takashi (Japan)
Fukumoto, Yoshihiro (Japan)
Fukushima, Masatoshi (Japan)
Furati, Khaled (Saudi Arabia)
Furusho, Hidekazu (Japan)
Furuta, Mikio (Japan)
Futorny, Vyacheslav (Brazil)
Gabasov, Rafail (Belarus)
Gachpazan, Morteza (Iran)
Gaitsgory, Dennis (USA)
Gajda, Wojciech (Poland)
Gamba, Irene M. (USA)

Gambarelli, Gianfranco (Italy)
Gamst, Jens (Germany)
Gan, Shaobo (China)
Gao, Enwei (China)
Gao, Hang (China)
Gao, Hongjun (China)
Gao, Hongzhu (China)
Gao, Ji (USA)
Gao, Jinwu (China)
Gao, Mingzhe (China)
Gao, Ran (China)
Gao, Shaogen (China)
Gao, Shouping (China)
Gao, Shu (China)
Gao, Song (China)
Gao, Ting (China)
Gao, Wenjie (China)
Gao, Xia (China)
Gao, Xiaoshan (China)
Gao, Ying (China)
Gao, Ying (Hubei) (China)
Gao, Yinghui (China)
Gao, Yuli (China)
Gao, Yun (Canada)
Gao, Zhicheng (Canada)
Gao, Zongsheng (China)
Gao, Zuofeng (China)
Gao, Zuxin (China)
Garcia Pupo, Mauro M. (Cuba)
Garcia Ruvalcaba, Jose de Jesus (Germany)
Gardiner, Frederick P. (USA)
Garling, David J. H. (United Kingdom)
Garner, Latonya C. (USA)
Garnica, Eugenio V. (Mexico)
Gasanov, Elyar E. (Russia)
Gasparyan, Armenak (Russia)
Gastaldi, Lucia (Italy)
Gatto, Eduardo (USA)
Gauthier, Paul Montpetit (Canada)
Gauthier, Yvon (Canada)
Ge, Jian (China)
Ge, Liming (USA)
Ge, Molin (China)
Ge, Weigao (China)
Ge, Yubo (China)
Ge, Yuxin (France)
Geddes, Keith O. (Canada)
Geiges, Hansjörg C. (Germany)
Geisser, Thomas H. (Japan)
Gelfand, Sergei (USA)
Geng, Jiansheng (China)
Geng, Xianmin (China)
Geraci, Linda L. (USA)
Gerasin, Sergey Nikolaevich (Ukraine)
Geroldinger, Alfred (Austria)
Gerstenkorn, Tadeusz (Poland)
Ghahramani, Saeed (USA)
Ghate, Eknath (India)
Ghisa, Dorin (Canada)
Ghorpade, Sudhir (India)
Ghosh, Shamindra Kumar (India)
Gibson, Paul (USA)
Gigena, Salvador D. R. (Argentina)
Gillard, Roland (France)
Giroux, Emmanuel (France)
Gitik, Moti (Israel)
Gloden, Raoul F. (Luxembourg)
Glowinski, Roland (USA)
Gne, Ver (France)
Godjali, Ali (USA)
Goetze, Friedrich (Germany)
Golan, Jonathan S. (Israel)
Golchin, Akbar (Iran)
Goldblatt, Robert (New Zealand)
Goldman, Irina E. (Russia)
Goldstein, Stanislaw (Poland)
Goldwasser, Shafi (USA)
Golov, Igor (Russia)
Golubeva, Valentina (Russia)
Golubyatnikov, Vladimir Petrovich (Russia)
Gomez, Xavier (United Kingdom)
Gomez Bermudez, Carlos (Spain)
Gomez Larranaga, Jose Carlos (Mexico)
Gomi, Kiyonori (Japan)
Goncalves , Daciberg Lima (Brazil)
Gondard, Danielle J. (France)
Gong, Fuzhou (China)
Gong, Guihua (USA)

Gong, Jianhua (New Zealand)
Gong, Ke (China)
Gong, Xunhua (China)
Gontcharov, Serguei Savostyanovich (Russia)
Gonzalez, Cesareo J. (Spain)
Gorbatsevich, Vladimir V. (Russia)
Gorbounov, Vassily (USA)
Gordon, Cameron M. (USA)
Gorjizadeh, Vali (Iran)
Gorkavyy, Vasyl A. (Ukraine)
Gorobets, Alexander D. (Ukraine)
Gorodski, Claudio (Brazil)
Gossez, Jean-Pierre (Belgium)
Goto, Midori (Japan)
Gottschalk, Hanno (Germany)
Göttsche, Lothar (Italy)
Gould, Mark (Australia)
Goulnara, Arjantseva (Russia)
Gowers, William T. (United Kingdom)
Gowrisankaran, Kohur (Canada)
Grafakos, Loukas (USA)
Graham, Ian R. (Canada)
Gramsch, Bernhard F. (Germany)
Grandsard, Francine (Belgium)
Grebenev, Vladimir (Russia)
Greenman, Jonathan V. (United Kingdom)
Greshnov, Alexander V. (Russia)
Greuel, Gert-Martin (Germany)
Griffiths, Phillip A. (USA)
Gritzmann, Peter (Germany)
Grobler, Jacobus J. (South Africa)
Groetschel, Martin (Germany)
Gromov, Nikolai Alekseevich (Russia)
Grossman, Daniel Andrew (USA)
Grunewald, Natalie (Germany)
Grushevsky, Samuel (USA)
Grushko, Pavel Jakovlevich (Russia)
Gryshko, Yuliya V. (Ukraine)
Grzymkowski, Radoslaw (Poland)
Gu, Alun (China)
Gu, Chaohao (China)
Gu, Chuanqing (China)
Gu, Guiding (China)
Gu, Sheng-Shi (China)
Gu, Weiqing (USA)
Gu, Yanling (China)
Gu, Yongxing (China)
Gu, Yundong (China)
Guan, Bo (USA)
Guan, Hua (China)
Guan, Qing Yang (China)
Guan, Yaji (China)
Guaraldo, Rosalind (USA)
Guest, Martin A. (Japan)
Guichardet, Alain (France)
Guillot, Pierre (United Kingdom)
Guo, Boling (China)
Guo, Dedian (China)
Guo, Jinyun (China)
Guo, Junwei (China)
Guo, Kui (China)
Guo, Kunyu (China)
Guo, Lei (China)
Guo, Maozheng (China)
Guo, Rui (China)
Guo, Shangjiang (China)
Guo, Shirong (China)
Guo, Shuguang (China)
Guo, Shunsheng (China)
Guo, Tie-Xin (China)
Guo, Wei (China)
Guo, Wenbin (China)
Guo, Xiaofeng (China)
Guo, Xiaohu (China)
Guo, Xuejun (China)
Guo, Yufeng (China)
Guo, Yuqi (China)
Guo, Zhiming (China)
Guo, Zongming (China)
Gupal, Anatolij Mihajlovych (Ukraine)
Gupta, Arjun (USA)
Gusev, Valery (Russia)
Gusev, Valery A. (Russia)
Guterman, Alexander E. (Russia)
Guzowska, Malgorzata (Poland)
Ha Huy, Khoai (Vietnam)
Haagerup, Uffe (Denmark)
Haase, Christian (USA)

Hachimori, Masahiro (Japan)
Hadid, Samir Bashir
(United Arab Emirates)
Hadwin, Donald W. (USA)
Haghani, Hosein Ahmad (Iran)
Hai, Phung Ho (Vietnam)
Hajarnavis, Charudatta R.
(United Kingdom)
Hajer, Bahouri (Tunesia)
Hajiabolhassan, Hossein (Iran)
Hales, Thomas C. (USA)
Haltar, Damba (Mongolia)
Hamana, Yuji (Japan)
Hamdari, Mehrdad (USA)
Han, Bin (Canada)
Han, Chong-Kyu (Korea)
Han, Danfu (China)
Han, Dong (China)
Han, Fei (China)
Han, Houde (China)
Han, Huili (China)
Han, Jinfang (China)
Han, Maoan (China)
Han, Pigong (China)
Han, Qing (China)
Han, Shijun (China)
Han, Xianglin (China)
Han, Xiaoli (China)
Han, Xinglou (China)
Han, Yang (China)
Han, Yanying (China)
Han, Yi Annie (USA)
Han, Yuliang (China)
Han, Zhangjia (China)
Hanada, Takao (Japan)
Hansen, Vagn Lundsgaard (Denmark)
Hao, Xuemei (China)
Hara, Masao (Japan)
Hara, Yuji (Japan)
Haran, Shai (Israel)
Haroutunian, Evgueni (Armenia)
Harris, Michael H. (France)
Hartz, David (USA)
Harvey, Shelly L. (USA)
Harzheim, Egbert R. (Germany)
Hasebe, Chino (Japan)
Hashemiparast, Seyed Moghtada (Iran)
Hassairi, Abdelhamid (Tunesia)
Hassoui, Yassine (Morocco)
Hasumi, Morisuke (Japan)
Haverhals, John S. (USA)
Haviar, Miroslav (Slovakia)
Hayakawa, Keizou (Japan)
He, Bin (China)
He, Chengchun (China)
He, Daojiang (China)
He, Guoqiang (China)
He, Hailong (China)
He, Hongyu (USA)
He, Hua (China)
He, Huixia (China)
He, Ji-Huan (China)
He, Jinchun (China)
He, Kun (China)
He, Lianfa (China)
He, Li-Ping (China)
He, Michael (USA)
He, Ping (China)
He, Pizhen (China)
He, Tian-Xiao (USA)
He, Wei (China)
He, Wenjie (China)
He, Wenming (China)
He, Xiaoyuan (Beijing)(China)
He, Xiaoyuan (China)
He, Xiqin (China)
He, Xisheng (USA)
He, Yijun (China)
He, Yongsen (China)
He, Yue (China)
He, Yuzan (China)
He, Zhen (China)
He, Zhenbang (China)
He, Zhendong (China)
He, Zhiqing (China)
He, Zuguo (China)
Hebert, Michel (Egypt)
Hege, Hans C. (Germany)
Heikkala, Ville (Finland)
Heinig, Hans P. (Canada)

Heinonen, Juha M. (USA)
Heinze, Joachim F. (Germany)
Hemakul, Wanida (Thailand)
Hemchote, Patcharin (Thailand)
Hemmecke, Raymond (USA)
Hemmer, David J. (USA)
Hempfling, Thomas (Switzerland)
Henstridge, John D. (Australia)
Herrero Blanco, Paloma (Spain)
Herrmann, Norbert (Germany)
Hesselholt, Lars (USA)
Hidaka, Fumio (Japan)
Hidalgo, Ruben (Chile)
Hideo, Shimada (Japan)
Hill, Theodore P. (USA)
Hilsum, Michel (France)
Hinojosa Luna, Adrian Pablo (Brazil)
Hintermann, Thomas (Switzerland)
Hinz, Andreas M. (Germany)
Hirachi, Yutaka (Japan)
Hirahata, Hirotoshi (Japan)
Hiraide, Koichi (Japan)
Hoang, Phu Xuan (Vietnam)
Hodges, Wilfrid A. (United Kingdom)
Hodgson, Bernard R. (Canada)
Hoevenaars, Luuk K. (Netherlands)
Hohberger, Horst (Germany)
Holgate, David B. (South Africa)
Holme, Audun (Norway)
Holte, John M. (USA)
Honary, Bahman (Iran)
Honda, Naofumi (Japan)
Hong, Bo (China)
Hong, Jialin (China)
Hong, Jiaxing (China)
Hong, Min-Chun (Australia)
Hopkins, Michael J. (USA)
Hora, Akihito (Japan)
Hori, Kentaro (USA)
Horita, Vanderlei Minori (Brazil)
Horiuchi, Toshio (Japan)
Hossein, Zand (United Kingdom)
Hou, Bo (China)
Hou, Cheng Jun (China)
Hou, Guozhen (China)
Hou, Jiangxia (China)
Hou, Jinchuan (China)
Hou, Roger H. (USA)
Hou, Shaosheng (China)
Hou, Shengzhao (China)
Hou, Xuezhang (USA)
Hou, Yumei (China)
Hou, Zhenting (China)
Hou, Zhong Hua (China)
Hou, Zixin (China)
Hou, Xiaoshan (China)
Houston, Johnny L. (USA)
Hoyles, Celia M. (United Kingdom)
Hrbacek, Karel (USA)
Hsiao, Ling (China)
Hsu, Pao-Sheng (USA)
Hu, Aiping (China)
Hu, Anli (China)
Hu, Ching-Lin (China)
Hu, Chuan-Gan (China)
Hu, Dihe (China)
Hu, Fannu (China)
Hu, Guoding (China)
Hu, Hesheng (China)
Hu, Jianhua (China)
Hu, Jianxun (China)
Hu, Jun (China)
Hu, Junfeng (China)
Hu, Naihong (China)
Hu, Peichu (China)
Hu, Po (USA)
Hu, Qiya (China)
Hu, Sen (China)
Hu, Shouchuan (USA)
Hu, Shuo (China)
Hu, Xian Fang (China)
Hu, Xiao-Dong (China)
Hu, Xiaohong (China)
Hu, Xijian (China)
Hu, Xijun (China)
Hu, Xingbiao (China)
Hu, Yi (China)
Hu, Yijun (China)
Hu, Yong (China)
Hu, Yueyun (France)

Hu, Zechun (China)
Hu, Zhangjian (China)
Huan, Zhongdan (China)
Huang, Anmin (China)
Huang, Cheng-Gui (China)
Huang, Chengming (Belgium)
Huang, Debin (China)
Huang, Hong (China)
Huang, Hongxuan (China)
Huang, Hua Lin (China)
Huang, Huaxiong (Canada)
Huang, Hui (China)
Huang, Huichi (China)
Huang, Jianguo (China)
Huang, Jianhua (China)
Huang, Jing (Canada)
Huang, Jingfang (USA)
Huang, Lihong (China)
Huang, Min (China)
Huang, Qianglian (China)
Huang, Qiong Xiang (China)
Huang, Tayuan (Taiwan)
Huang, Weiming (China)
Huang, Wen (China)
Huang, Wen-Ling (Germany)
Huang, Xiaojun (China)
Huang, Xiaoling (China)
Huang, Xinzhong (China)
Huang, Yiru (China)
Huang, Yong (China)
Huang, Yonglong (China)
Huang, Yuanfei (China)
Huang, Yuanqiu (China)
Huang, Yongwei (China)
Huang, Zhanyou (China)
Huang, Zhaobo (China)
Huang, Zhaohui (China)
Huang, Zhaoyong (China)
Huang, Zhengda (China)
Huang, Zhihong (China)
Huang, Zhiyong (China)
Huang, Zhong Lin (China)
Huang, Zhongyi (China)
Huber, Annette (Germany)
Hughes, Anne C. (USA)
Hughes, Kenneth R. (South Africa)
Hui, Changchang (China)
Hui, Jing (China)
Huisman, Johannes (France)
Hunt, David C. (Australia)
Hurtubise, Jacques C. (Canada)
Husemöller, Dale H. (Germany)
Ibrahim, Slim (Italy)
Igarashi, Masayuki (Japan)
Iitaka, Shigeru S. (Japan)
Ikeda, Satoshi (Japan)
Ikeda, Takeshi (Japan)
Illusie, Luc (France)
Ilolov, Mamadsho (Tajikistan)
Im, Bokhee (Kyrgyzstan)
Imanaliev, Myrzabek Imanalievich (Kyrgyzstan)
Impagliazzo, Russell Graham (USA)
Inassaridze, Nick (Georgia)
Infante, Gennaro (Italy)
Inprasit, Utith (Thailand)
Inshakov, Andrey (Russia)
Ion, Patrick D. F. (USA)
Ionel, Eleny-Nicoleta (USA)
Ionescu, Cristodor H. (Romania)
Iorio Jr., Rafael J. (Brazil)
Ioudine, Vladimir (Russia)
Iranmanesh, Ali (Iran)
Ishikawa, Takeshi (Japan)
Iskhokov, Sulaimon Abunasrovich (Tajikistan)
Ito, Yoshihiko (Japan)
Ito, Yukari (Japan)
Itoh, Jin-Ichi (Japan)
Iusem, Alfredo N. (Brazil)
Izmestiev, Ivan (Germany)
Izumori, Hitoshi (Japan)
Jacimovic, Milojica (Yugoslavia)
Jackson, Allyn (Germany)
Jacob, Mostovoy (Russia)
Jacobowitz, Howard (USA)
Jahnke, Hans Niels (Germany)
Janphaisaeng, Suphawan (Thailand)
Jeltsch, Rolf (Switzerland)
Jeon, Intae (Korea)

Jeong, Chan-Seok (Korea)
Ji, Bo (China)
Ji, Chungang (China)
Ji, Guanghua (China)
Ji, Guangzhong (China)
Ji, Lizhen (USA)
Ji, Min (China)
Ji, Xia (China)
Ji, Xinhua (China)
Ji, You Qing (China)
Ji, Yuan (China)
Ji, Yue (China)
Jia, Chaohua (China)
Jia, Fang (China)
Jia, Houyu (China)
Jia, Yueling (China)
Jian, Huaiyu (China)
Jiang, Boju (China)
Jiang, Chengshun (China)
Jiang, Chunlan (China)
Jiang, Cuibo (China)
Jiang, Dihua (USA)
Jiang, Erxiong (China)
Jiang, Jian (China)
Jiang, Jiancheng (China)
Jiang, Jiao (China)
Jiang, Lining (China)
Jiang, Lishang (China)
Jiang, Mei-Yue (China)
Jiang, Miao (China)
Jiang, Murong (China)
Jiang, Song (China)
Jiang, Tiantian (China)
Jiang, Tiejun (China)
Jiang, Wei (China)
Jiang, Xingyao (China)
Jiang, Yinsheng (China)
Jiang, Yiwen (China)
Jiang, Yunping (USA)
Jiang, Zhimin (China)
Jiao, Baocong (China)
Jiao, Rongzheng (China)
Jiao, Xiaoxiang (China)
Jiao, Yang (China)
Jiao, Zhengming (China)
Jimenez, Miguel Antonio (Cuba)
Jin, Chunyan (China)
Jin, Congming (China)
Jin, Gan (China)
Jin, Guitang (China)
Jin, Jiashun (USA)
Jin, Mingzhong (China)
Jin, Xiao-Qing (China)
Jin, Yafen (China)
Jin, Yang (China)
Jin, Yinghua (Korea)
Jin, Zheng Guo (Korea)
Jing, Guangting (China)
Jing, Jian (China)
Jing, Xiao Ying (China)
Jing, Zhujun (China)
Jinzenji, Masao (Japan)
Jitomirskaya, Svetlana (USA)
Jo, Jang Hyun (Korea)
Joe, Kamaimoto (Japan)
Joffe, Anatole (Canada)
Johansson, Kurt (Sweden)
Jones, Alison M. (United Kingdom)
Jong, Jae-Bu (DPR of Korea)
Jong, Kwang Ho (DPR of Korea)
Joshi, Mohan C. (India)
Joswig, Michael (Germany)
Jovanovic, Bosko S. (Yugoslavia)
Joyce, Dominic D. (United Kingdom)
Jozefiak, Tadeusz (USA)
Ju, Hyeong-Kwan (Korea)
Ju, Qiangchang (China)
Juajun, Jeeranunt (Thailand)
Jugjai, Srisangiem (Thailand)
Juntharee, Pongpol (Thailand)
Juriaans, Stanley Orlando (Brazil)
Juutinen, Petri A. (Finland)
Kabeya, Yoshitsugu (Japan)
Kabiljanova, Firuza (Uzbekistan)
Kac, Victor (USA)
Kadison, Richard V. (USA)
Kaenmaki, Antti (Finland)
Kahn, Bruno (France)
Kaijser, Sten G. (Sweden)
Kaino, Luckson M. (Botswana)

Kaiser, Gabriele (Germany)
Kaji, Hajime (Japan)
Kaji, Yamamuro (Japan)
Kajimoto, Hiroshi (Japan)
Kakishov, Kanybek Kakishovich (Kyrgyzstan)
Kallunki, Sari A. (Finland)
Kalnitsky, Vyacheslav S. (Russia)
Kalyabin, Gennadiy (Russia)
Kamei, Tetsujiro (Japan)
Kan, Jiahai (China)
Kan, Rui (China)
Kanarek, Herbert (Mexico)
Kandaswamy, Palani Gounder (India)
Kaneko, Makoto (Japan)
Kaneyuki, Soji (Japan)
Kang, Qingde (China)
Kang, Yifang (China)
Kangro, Urve (Estonia)
Kannan, Ravi (USA)
Kaplan, Jonathan R. (USA)
Kappos, Efthimios (United Kingdom)
Kapur, Aruna (India)
Kara, Abdul Hamid (South Africa)
Karabudak, Ersin (Turkey)
Karailiev, Lubomir Alexandrov (Russia)
Karamzadeh, O.A.S. (Iran)
Karandikar, Rajeeva L. (India)
Karavan Jahromi, Mehrdad (Iran)
Kasahara, Yasushi (Japan)
Kashani, S.M.B. (Iran)
Kassel, Christian (France)
Kasyanov, Victor (Russia)
Katayama, Shinichi (Japan)
Kathotia, Vinay (China)
Kato, Kazuya (Japan)
Katona, Gyula (Hungary)
Katre, Shashikant Anant (India)
Katsap, Ada (Israel)
Katsura, Takeshi (Japan)
Katz, Alexander A. (USA)
Kaufmann, Ralph M. (Germany)
Kawano, Nichiro N.K. (Japan)
Kazakov, Vladimir A. (Mexico)
Ke, Zhao (China)
Keedwell, A.D. (United Kingdom)
Keeton, Stephine L. (USA)
Keller, Bernhard (France)
Keller, Thomas Michael (USA)
Kemp, Paula A. (USA)
Kenig, Carlos E. (USA)
Keppelmann, Edward (USA)
Kerayechian, Asghar (Iran)
Kerman, Ron A. (Canada)
Kesemen, Tulay (Turkey)
Kesten, Harry (USA)
Ketchanwit, Pasakorn (Thailand)
Keum, Jonghae (Korea)
Keune, Frans J. (Netherlands)
Khabeev, Nail Suleiman (Bahrain)
Khaldjigitov, Abduvaly (Uzbekistan)
Khan, Qamar Jalil Ahmad (Oman)
Khaniyev, Tahir (Turkey)
Kharchenko, Vladislav (Mexico)
Khechinashvili, Zaza (Georgia)
Kheifets, Igor L. (Russia)
Khmaladze, Emzar (Georgia)
Khoa Son, Nguyen (Vietnam)
Khongtham, Yaowaluck (Thailand)
Khoshkam, Mahmood (Canada)
Khosla, Deepee (USA)
Khosrovshahi, Gholamreza B. (Iran)
Khots, Boris (USA)
Khots, Dmitriy (USA)
Khumsup, Preya (Thailand)
Kiguradze, Ivan (Georgia)
Kiguradze, Tariel (Georgia)
Kikuchi, Kazunori (Japan)
Kilpelainen, Tero (Finland)
Kim, Ae-Shim (Korea)
Kim, Daeyong (Korea)
Kim, Dohan (Korea)
Kim, Dong-Soo (Korea)
Kim, Eui-Chul (Korea)
Kim, Gol (DPR of Korea)
Kim, Hoil (Korea)
Kim, Hyuk (Korea)
Kim, Il Ho (Korea)
Kim, Inkang (Korea)
Kim, Jeong-Hoon (Korea)

Kim, Jin-Hong (Korea)
Kim, Jongsu (Korea)
Kim, Joonhyung (Korea)
Kim, Kwang I. (Korea)
Kim, Mun Chol (DPR of Korea)
Kim, Myung-Hwan (Korea)
Kim, Seung Won (Korea)
Kim, Sun-Chul (Korea)
Kim, Tu Jin (Korea)
Kim, Wan S. (Korea)
Kim, Young Ho (Korea)
Kimura, Noriaki (Japan)
Kimura, Takashi (USA)
Kiranagi, Basavannappa S. (India)
Kireitseu, Maksim (Belarus)
Kirillov, Oleg Nikolayevich (Russia)
Kirillova, Faina Mikhailovna (Belarus)
Kirschenhofer, Peter (Austria)
Kisaka, Masashi (Japan)
Kiselman, Christer O. (Sweden)
Kisilevsky, Hershy H. (Canada)
Kister, James M. (USA)
Kister, Jane E. (USA)
Kitaev, Alexei Y. (USA)
Kitano, Teruaki (Japan)
Kitazaki, Kuniaki (Japan)
Kitover, Arkady K. (USA)
Kiwi, Jan (Chile)
Kiyohara, Kazuyoshi (Japan)
Kkawano, Nichiro (Japan)
Klecha, Tadeusz (Poland)
Kleive, Per-Even (Norway)
Klodginski, Elizabeth A. (USA)
Kluza, Krzysztof (Poland)
Klyachko, Alexander (Turkey)
Knezevic-Miljanovic, Julka (Yugoslavia)
Knight, Robert (USA)
Ko, Bong Soo (Korea)
Ko, Hyoung J. (Korea)
Ko, Younhee (Korea)
Kobayashi, Toshiyuki (Japan)
Kochubei, Anatoly N. (Ukraine)
Kodama, Masao (Japan)
Kodiyalam, Vijay (India)
Koelblen, Laurent (France)
Koenig, Gerhard (Germany)
Kohn, Robert V. (USA)
Kojima, Hisashi (Japan)
Kojok, Badrié (Lebanon)
Kokilashvili, Vakhtang (Georgia)
Kokubu, Hiroshi (Japan)
Kolesnikov, Pavel (Russia)
Komatsu, Hikosaburo (Japan)
Kon, Benzion A. (Israel)
Kondo, Takefurni K. (Japan)
Kondratiev, Vladimir (Russia)
Kondratieva, Margarita (USA)
Kong, De-Xing (China)
Kong, Hui (China)
Kononenko, Larisa Ivanovna (Russia)
Konstantinova, Elena (Russia)
Kooburat, Paktra (Thailand)
Kopanskii, Alexander (Moldova)
Kopell, Nancy Jane (USA)
Koranyi, Adam (USA)
Korhonen, Risto J. (Finland)
Korobitsin, Victor (Russia)
Kosinski, Antoni A. (USA)
Koskela, Pekka J. (Finland)
Kota, Osamu (Japan)
Kotschick, Dieter (Germany)
Kou, Hui (China)
Kou, Kit Ian (China)
Koudriavtseva, Elena (Russia)
Kouniavski (Kunyavskii), Boris (Israel)
Kouzoub, Natalie M. (Russia)
Kovacs, Sandor (USA)
Kovalevsky, Alexander Albertovich (Ukraine)
Kowalski, Piotr L. (Poland)
Koyama, Shinya (Japan)
Kozhevnikov, Alexander (Israel)
Kozlovski, Oleg (United Kingdom)
Kraljevic, Hrvoje (Croatia)
Krantz, Steven G. (USA)
Kranz, Przemo T. (USA)
Krason, Piotr (Poland)
Krieger, Wolfgang J. (Germany)
Krilic-Smajlovic, Lejla (Bosnia and Herzegovina)

Krishnan, Balachandran (India)
Krishnan, Edamana Vasudevan (Oman)
Kriz, Igor (USA)
Kroger, Heinz (Germany)
Kruglikov, Boris S. (Russia)
Krukier, Lev A. (Russia)
Krutitskii, Pavel (Russia)
Ku, Chul (Korea)
Kubarski, Jan (Poland)
Kucherenko, Valeri (Mexico)
Kucuk, Zafer (Turkey)
Kudla, Stephen S. (USA)
Kudryashov, Nikolai A. (Russia)
Kufner, Alois (Czech)
Kuku, Aderemi O. (Italy)
Kumar, Sanjeev (India)
Kumar, Vinod (India)
Kunzle, Hans-Peter (Canada)
Kurganov, Alexander (USA)
Kusuoka, Shigeo (Japan)
Kutyshkin, Andrey Valentinovich (Russia)
Kutzschebauch, Frank (Sweden)
Kuznetsov, Vadim B. (United Kingdom)
Kwon, Doyong (Korea)
Labahn, George (Canada)
Labarca, Rafael E. (Chile)
Lackenby, Marc (United Kingdom)
Lafforgue, Laurent (France)
Lafforgue, Vincent G. (France)
Lai, Ah Moey (Malaysia)
Lai, Chunhui (China)
Lai, Wancai (China)
Lai, Yim Fun (Malaysia)
Lakhtin, Alexey S. (Russia)
Lam, Kee Yuen (Canada)
Lan, Shu (China)
Lan, Yizhong (China)
Lance, Christopher (United Kingdom)
Lange, Carsten (Germany)
Lap, James T. (USA)
Lapid, Erez (USA)
Lapidus, Michel L. (USA)
Laptev, Gennady G. (Russia)
Laptev, Gennady I. (Russia)
Lascar, Daniel (France)
Lashkhi, Alexander A. (Georgia)
Latala, Rafal (Poland)
Latipov, Halim Rafikovich (Uzbekistan)
Laudal, Olav A. (Norway)
Laumon, Géard (France)
Laurincikas, Antanas (Lithuania)
Lavrauw, Michel (Italy)
Lawler, Gregory Francis (USA)
Le, Ha Quang (Canada)
Le Roux, Jonathan (France)
Le Thi, Hoai An (France)
Le Thi Thanh, Nhan (Vietnam)
Lee, Chul Woo (Korea)
Lee, Dongyoung (Korea)
Lee, Guiqing (China)
Lee, Hyun-Dae (Korea)
Lee, Jong Bum (Korea)
Lee, Keonhee (Korea)
Lee, Nam Kee (Korea)
Lee, Richard Wm (Canada)
Lee, Sang Min (Korea)
Lee, Sang Youl (Korea)
Lee, Seung Won (Korea)
Lee, Sungchul (Korea)
Lee, Tsiu-Kwen (Taiwan)
Lee, Yongnam (Korea)
Lee, Yoonweon (Korea)
Lee, Young Joo (Korea)
Lee, Yuan-Pin (USA)
Lee, Yuwen (China)
Leerawat, Utsanee (Thailand)
Legkin, Anton (USA)
Lei, Fengchun (China)
Lei, Jianguo (China)
Lei, Jinzhi (China)
Lei, Lee (China)
Lei, Siu-Long (China)
Lei, Tiangang (China)
Lei, Xubo (China)
Lei, Zheng (China)
Leichtweiss, Kurt (Germany)
Leitao, Antonio C. (Brazil)
Lemmermeyer, Franz (USA)
Lempert, Laszlo (USA)

Lenci, Marco (USA)
Leng, Gangsong (China)
Leng, Wilfrid S. (United Kingdom)
Lerman, Lev (Russia)
Lerner, Nicolas (France)
Lesmes, Jaime Ignacio (Colombia)
Lester, June A. (Canada)
Leung, Frederick Koon-Shing (China)
Levesque, Claude (Canada)
Levine, Marc N. (USA)
Levy, Doron (USA)
Lewkeeratiyutkul, Wicharn (Thailand)
Leykin, Anton (USA)
Li, Aihua (USA)
Li, Angsheng (China)
Li, Anmin (China)
Li, Banghe (China)
Li, Baokui (China)
Li, Baoyuan (China)
Li, Bing (China)
Li, Bingren (China)
Li, Bingzhao (China)
Li, Bingzheng (China)
Li, Bo (China)
Li, Changguo (China)
Li, Changjun (Japan)
Li, Changpin (China)
Li, Chao (China)
Li, Charles (USA)
Li, Chengyue (China)
Li, Chengzhi (China)
Li, Chong (China)
Li, Congzhu (China)
Li, Cuixiang (China)
Li, Dafa (China)
Li, Daqian (China)
Li, Delang (China)
Li, Dengxin (China)
Li, Deqin (China)
Li, Desheng (China)
Li, Dong (China)
Li, Dongsheng (China)
Li, Duoquan (China)
Li, Fang (China)
Li, Fei (China)
Li, Fuan (China)
Li, Gaidi (China)
Li, Gang (China)
Li, Gongbao (China)
Li, Gongnong (China)
Li, Gongsheng (China)
Li, Guiqing (China)
Li, Guoying (China)
Li, Haidong (China)
Li, Hailiang (Japan)
Li, Haitao (China)
Li, Hong (China)
Li, Hongbo (China)
Li, Hongwei (China)
Li, Hongxing (China)
Li, Hongze (China)
Li, Hui (China)
Li, Huiling (China)
Li, Huiyuan (China)
Li, Jia (China)
Li, Jian (China)
Li, Jiangfeng (China)
Li, Jiangtao (China)
Li, Jianping (China)
Li, Jian-Shu (China)
Li, Jiasheng (China)
Li, Jiayong (China)
Li, Jiayu (China)
Li, Jie (China)
Li, Jiequan (Germany)
Li, Jing-An (China)
Li, Jinlu (USA)
Li, Juan (China)
Li, Jun (China)
Li, Kedian (China)
Li, Kezheng (China)
Li, Kuiyuan (USA)
Li, Lei (China)
Li, Liang (China)
Li, Libin (China)
Li, Lili (China)
Li, Linsong (Korea)
Li, Longcai (China)
Li, Luoqing (China)
Li, Min (China)

Li, Ming (China)
Li, Minli (China)
Li, Na (China)
Li, Peixin (China)
Li, Peter W. (USA)
Li, Qiao (China)
Li, Qihui (China)
Li, Qingguo (China)
Li, Qingzhong (China)
Li, Qiong (China)
Li, Qisheng (China)
Li, Shangzhi (China)
Li, Shenghong (China)
Li, Shengjia (China)
Li, Shoumei (China)
Li, Shujie (China)
Li, Tianhong (USA)
Li, Tianze (China)
Li, Tiecheng (China)
Li, Tiejun (China)
Li, Tong (USA)
Li, Wan-Tong (China)
Li, Wei (China)
Li, Wei (AMSS) (China)
Li, Wei (Tsinghua) (China)
Li, Weimin (China)
Li, Weinian (China)
Li, Wei-Ping (China)
Li, Weiqing (China)
Li, Wenbo V. (USA)
Li, Wenlin (China)
Li, Wenlin (Henan) (China)
Li, Wen-Ming (China)
Li, Xia (China)
Li, Xiang Dong (United Kingdom)
Li, Xiangyang (China)
Li, Xianhua (China)
Li, Xianliang (China)
Li, Xiaodong (China)
Li, Xiaodong (USA)
Li, Xiaoyue (China)
Li, Xingxiao (China)
Li, Xinmin (China)
Li, Xiong (China)
Li, Xueliang (China)
Li, Xuemei (China)
Li, Xuewen (China)
Li, Xuhong (China)
Li, Yachun (China)
Li, Yangcheng (China)
Li, Yangming (China)
Li, Yangrong (China)
Li, Yanyan (USA)
Li, Yaotang (China)
Li, Yaqing (China)
Li, Yezhou (China)
Li, Yong (China)
Li, Yong (Jilin) (China)
Li, Yonghai (China)
Li, Yongkun (China)
Li, Yongsheng (China)
Li, Youai (China)
Li, Youyun (China)
Li, Yuan (China)
Li, Yuanlin (Canada)
Li, Yuhua (China)
Li, Yusheng (China)
Li, Yuwen (China)
Li, Yuxiang (China)
Li, Zemin (China)
Li, Zenghu (China)
Li, Zhenguo (China)
Li, Zhenping (China)
Li, Zhenqi (China)
Li, Zhiwei (China)
Li, Zhong (China)
Li, Zhongkai (China)
Li, Zhongyan (China)
Lian, Qiaofang (China)
Liang, Chao (China)
Liang, Degang (China)
Liang, Dengfeng (China)
Liang, Guoping (China)
Liang, Heng (China)
Liang, Hengyi (China)
Liang, Jihua (China)
Liang, Ke (China)
Liang, Margaret (Canada)
Liang, Song (Japan)
Liang, Xian (China)

Liang, Xinyuan (China)
Liang, Xue-Zhang (China)
Liang, Yawei (Canada)
Liang, Ye (China)
Liang, Zhibin (China)
Liang, Zongxia (China)
Liao, Caisheng (China)
Liao, Fucheng (China)
Liao, Ling Min (China)
Liao, Liusheng (China)
Liao, Moxiang (China)
Liao, Zuhua (China)
Liberzon, Mark (Russia)
Lichtenberg, Heiner (Germany)
Liczberski, Piotr (Poland)
Lie, Chang Hoon (Korea)
Lie, Wei (China)
Lim, Chaeho (Korea)
Lim, Chong-Keang (Malaysia)
Lima, Eduardo (Venezuela)
Lima, Suely (Brazil)
Lin, Dongdai (China)
Lin, Dongyu (China)
Lin, Fanghua (USA)
Lin, Guanghua (China)
Lin, Guanjun (China)
Lin, Hai Xiang (Netherlands)
Lin, Huazhen (China)
Lin, Jeng-Eng (USA)
Lin, Jiafu (China)
Lin, Jianrong (China)
Lin, Lei (China)
Lin, Liangyu (China)
Lin, Lijun (China)
Lin, Miao (China)
Lin, Ping (China)
Lin, Qun (China)
Lin, Qun (Beijing) (China)
Lin, Wei (China)
Lin, Xiaoping (China)
Lin, Xiaotao (China)
Lin, Yanping (China)
Lin, Yong (China)
Lin, Yuanlie (China)
Lin, Zhengbao (China)
Lin, Zhengyan (China)
Lin, Zongzhu (USA)
Lindemann, Ina (USA)
Lindstrom, Torsten A. (Sweden)
Ling, Guo-ping (China)
Linial, Nathan (Israel)
Linton, Fred E.J. (USA)
Lipachev, Evgeny (Russia)
Lipponen, Marjo R. (Finland)
Lipschutz, Seymour (USA)
Liseikin, Vladimir D. (Russia)
Liu, Baokang (China)
Liu, Bin (China)
Liu, Bing (China)
Liu, Biyue (USA)
Liu, Changmao (China)
Liu, Chao (China)
Liu, Chen (China)
Liu, Chungen (China)
Liu, Chunping (China)
Liu, De (China)
Liu, Feng (China)
Liu, Fengmei (China)
Liu, Fengshan (USA)
Liu, Fengsui (China)
Liu, Fon Che (China)
Liu, Gang (China)
Liu, Geng (China)
Liu, Genqian (China)
Liu, Guiju (China)
Liu, Guoxin (China)
Liu, Guoyang (USA)
Liu, Hailiang (USA)
Liu, Haitao (China)
Liu, Hao (China)
Liu, Heguo (China)
Liu, Heping (China)
Liu, Hongwei (China)
Liu, Huanping (China)
Liu, Huarong (China)
Liu, Hui (China)
Liu, Huizhao (China)
Liu, Jianguo (USA)
Liu, Jianming (China)
Liu, Jianya (China)

Liu, Jinlin (China)
Liu, Jun (China)
Liu, Junhong (China)
Liu, Juxin (China)
Liu, Kangsheng (China)
Liu, Ke (China)
Liu, Kefeng (USA)
Liu, Leping (China)
Liu, Li (China)
Liu, Liang (China)
Liu, Lianqi (China)
Liu, Lianxu (China)
Liu, Lifang (China)
Liu, Lijun (China)
Liu, Limin (China)
Liu, Liwei (China)
Liu, Lixin (China)
Liu, Mingju (China)
Liu, Mulan (China)
Liu, Ning (China)
Liu, Peide (China)
Liu, Peidong (China)
Liu, Qihao (China)
Liu, Qingping (China)
Liu, Qingyue (China)
Liu, Quansheng (France)
Liu, Runqiu (China)
Liu, Shangping (China)
Liu, Shaowu (China)
Liu, Sheng-Qiang (China)
Liu, Shi Zhu (China)
Liu, Shibo (China)
Liu, Shilian (China)
Liu, Shiqiang (China)
Liu, Tai-Ping (Taiwan)
Liu, Tianhui (China)
Liu, Tiantian (China)
Liu, Tingting (China)
Liu, Wanmin (China)
Liu, Weijun (China)
Liu, Weiming (China)
Liu, Wen (China)
Liu, Wenan (China)
Liu, Wenjun (China)
Liu, Winbin (China)
Liu, Xiangguan (China)
Liu, Xianning (China)
Liu, Xianquan (China)
Liu, Xiaobo (USA)
Liu, Xiaochuang (China)
Liu, Xiaoyan (USA)
Liu, Xingdong (China)
Liu, Xinsheng (China)
Liu, Xinzhi (Canada)
Liu, Xuefei (China)
Liu, Xuezhe (China)
Liu, Xufeng (China)
Liu, Yanchun (China)
Liu, Yian-Kui (China)
Liu, Ying (China)
Liu, Yingming (China)
Liu, Yong (China)
Liu, Yongmin (China)
Liu, Yongping (China)
Liu, Yongqin (China)
Liu, Youming (China)
Liu, Yuan (China)
Liu, Yuanzhang (China)
Liu, Yurong (China)
Liu, Zaiming (China)
Liu, Zengfan (China)
Liu, Zhan (China)
Liu, Zhangju (China)
Liu, Zhaoxia (China)
Liu, Zhicong (China)
Liu, Zhihua (China)
Liu, Zhongkui (China)
Liu, Zhuxing (China)
Liu, Zihui (China)
Liu, Ziyi (China)
Liu, Zuhan (China)
Liverpool, Lennox S. O. (Nigeria)
Lloyd, Edward Keith (United Kingdom)
Long, Guo Gui (China)
Long, Quan (China)
Long, Yiming (China)
Long, Zhengwu (China)
Longani, Vites (Thailand)
Longhi, Ignazio (Italy)
Lope, Jose Ernie Capioso (Philippines)

Lorenz, Falko (Germany)
Lou, Jie (China)
Lou, Yuanbing (China)
Lou, Zengjian (Australia)
Lozier, Daniel W. (USA)
Lu, Caihui (China)
Lu, Chungui (China)
Lu, Diming (China)
Lu, Guangcun (China)
Lu, Guofu (China)
Lu, Guozhen (USA)
Lu, Heng (China)
Lu, Hong (China)
Lu, Hongwen (China)
Lu, Hongying (China)
Lu, Jinhu (China)
Lu, Keping (China)
Lu, Linzhang (China)
Lu, Mengwen (China)
Lu, Minggao (China)
Lu, Ning (China)
Lu, Peili (China)
Lu, Qikeng (China)
Lu, Qishao (China)
Lu, Ruqian (China)
Lu, Shannian (China)
Lu, Shanzhen (China)
Lu, Shijie (China)
Lu, Shujuan (China)
Lu, Sishun (China)
Lu, Songsong (China)
Lu, Tiao (China)
Lu, Xuan (China)
Lu, Yao (China)
Lu, Ye-Guang (China)
Lu, Ying (China)
Lu, Yong (China)
Lu, Youmin (USA)
Lu, Yungang (Italy)
Lu, Yunguang (Colombia)
Lu, Zaiping (China)
Lu, Zhiqi (China)
Lu, Zhiqin (USA)
Lu, Zhonghua (China)
Lu, Zhujia (China)
Lü, Fang (China)
Lü, Feng (China)
Lü, Guo Yi (China)
Lü, Haishen (China)
Lü, Heng (China)
Lü, Jin-Hu (China)
Lü, Zhi (China)
Lü, Zhongxue (China)
Luan, Jing (China)
Luan, Jingwen (China)
Luan, Liangyu (China)
Luca, Florian (Mexico)
Ludwig, Garry (Canada)
Luetkebohmert, Werner (Germany)
Lun, Anthony W. (Australia)
Luna, Adrian Pablo Hinon (Brazil)
Luo, Haigang (China)
Luo, Jiaowan (China)
Luo, Maokang (China)
Luo, Qiu-Ming (China)
Luo, Shunlong (China)
Luo, Weidong (China)
Luo, Wenzhi (USA)
Luo, Xianshun (China)
Luo, Xiaohui (China)
Luo, Xuebo (China)
Luo, Yang (United Kingdom)
Luo, Yanning (China)
Luo, Yanxia (China)
Luo, Youfeng (China)
Luo, Zhiguo (China)
Lusala, Tsasa (Germany)
Luskin, Mitchell B. (USA)
Lussardi, Luca (Italy)
Lutterodt, Clement H. (USA)
Lutz, Frank H. (Germany)
Lykova, Zinaida A. (United Kingdom)
Lyons, Terry (United Kingdom)
Ma, Chuangui (China)
Ma, Fuming (China)
Ma, Guanzhong (China)
Ma, Guoxuan (China)
Ma, Hongbin (China)
Ma, Hui (China)
Ma, Jinxi (China)

Ma, Li (China)
Ma, Lianrong (China)
Ma, Qi-Wei (China)
Ma, Renyi (China)
Ma, Ruiqin (China)
Ma, Sheng-Ming (China)
Ma, Shiwang (China)
Ma, Wanbiao (China)
Ma, Wancang (USA)
Ma, Xia (China)
Ma, Xiaonan (France)
Ma, Xuan (China)
Ma, Yujie (China)
Ma, Yumei (China)
Ma, Zhiming (China)
Ma, Zhongtai (China)
Mabizela, Sizwe G. (South Africa)
Macarini, Leonardo Magalhaes (Brazil)
Macedonska, Olga (Poland)
Maclagan, Diane M. (USA)
Madan, Shobha (India)
Madanshekaf, Ali (Iran)
Madsen, Ib H. (Denmark)
Maeng, Ju Ok (Korea)
Maffei, Andrea (Italy)
Mahmoodian, Ebadollah S. (Iran)
Maillot, Vincent O. (France)
Mailybaev, Alexei A. (Russia)
Maiybaev, Alexei (Russia)
Majumdar, Subrata (Bangladesh)
Makharadze, Shota (Georgia)
Makhnev, Alexandre A. (Russia)
Makienko, Peter M. (Mexico)
Makin, Alexander S. (Russia)
Makinde, Oluwole (South Africa)
Makinde, Oluwole Daniel (South Africa)
Malinin, Dmitry A. (Belarus)
Malmini, Ranasinghe P. K. C. (Sri Lanka)
Mamadaliev, N.K. (Uzbekistan)
Mampassi, Benjamin (Senegal)
Manav, Das (India)
Manchanda, Pammy (India)
Mandai, Takeshi (Japan)
Manderscheid, David C. (USA)
Mann, William R. (USA)
Manturov, Vassily Olegovich (Russia)
Manuilov, Vladimir M. (Russia)
Mao, Dekang (China)
Mao, Rui (China)
Mao, Yonghua (China)
Mao, Yongsheng (China)
Maqsood, Tariq (Pakistan)
Marathe, Kishore (USA)
Marchisio, Marina (Italy)
Marcolli, Matilde (Germany)
Maria, Contessa (Italy)
Marian, Alina (USA)
Maritz, Pieter (South Africa)
Mark, Thomas E. (USA)
Markarian, Roberto (Uruguay)
Marki, Laszlo (Hungary)
Markina, Irina (Chile)
Markovskaya, Natalia (Belarus)
Martio, Olli T. (Finland)
Maruyama, Fumitsuna (Japan)
Masjedjamei, Mohammad (Iran)
Mathieu, Martin (United Kingdom)
Matsubara, Kiyoshi (Japan)
Matsuda, Osamu (Japan)
Matsuhisa, Takashi (Japan)
Matsumoto, Shigenori (Japan)
Matsumoto, Yukio (Japan)
Matsuyama, Yoshio (Japan)
Matveev, Serguei Vladimirovich (Russia)
Matveev, Vladimir (Germany)
Matveeva, Inessa I. (Russia)
Maude, Ronald (United Kingdom)
Maumary, Serge (Switzerland)
Mazov, Bogdan (Russia)
Mazov, Bogdan L. (Russia)
Mazurov, Victor D. (Russia)
Maz'ya, Vladimir (Sweden)
Mcdonald, Bernard Robert (USA)
Mcgerty, Kevin Rory (USA)
Mcintosh, Alan G. (Australia)
Mcmurray, Nolan B. (USA)
Mcquillan, Michael Liam (France)
Mechenov, Alexander (Russia)

Medeiros, Nivaldo Nunes (Brazil)
Mehri, Bahman (Iran)
Mehta, Ghanshyam Bhagvandas (Australia)
Mehta, Vikram Bhagvandas (India)
Mei, Jia Qiang (China)
Mei, Jialiu (China)
Mei, Xiangming (China)
Meier, David (Switzerland)
Meinrenken, Eckhard (Canada)
Mendonca, Sergio Jose (Brazil)
Meng, Bin (China)
Meng, Daoji (China)
Meng, Fanwei (China)
Meng, Guangwu (China)
Meng, Jixiang (China)
Meng, Kaitao (China)
Meng, Qingxun (China)
Meng, Yan (China)
Mercedes, Fernandez G. (Spain)
Merino, Glicina (Mexico)
Meskhi, Alexander N. (Georgia)
Mestel, Benjamin D. (United Kingdom)
Metzger, Roger Javier (Peru)
Meyer, Johan H. (South Africa)
Miao, Long (Hefei)(China)
Miao, Long (China)
Miao, Zhengke (China)
Micheletti, Anna Maria (Italy)
Michor, Peter W. (Austria)
Michtchenko, Tatiana M. (Russia)
Mickelsson, Jouko A. (Sweden)
Micula, George (Romania)
Mielke, Alexander (Germany)
Mielke, Marvin V. (USA)
Miguel, Jose Joao (Mozambique)
Mijatovic, Aleksandar (United Kingdom)
Mikami, Toshio (Japan)
Milan, Francisco (Spain)
Milanes, Aniura (Brazil)
Milka, Anatoliy D. (Ukraine)
Miller, Richard R. (USA)
Millett, Kenneth C. (USA)
Millionschikov, Dmitri (Russia)
Milnes, Paul (Canada)
Mimouni, Abdeslam (Morocco)
Min, Lequan (China)
Ming, Pingbing (China)
Minn, Jooha (Korea)
Mirakhmedov, Sherzod (Uzbekistan)
Miranda, Annamaria (Italy)
Miranville, Alain M. (France)
Misaki, Norihiro (Japan)
Mishchenko, Alexandr S. (Russia)
Mishura, Yuliya (Ukraine)
Misra, Kailash C. (USA)
Mitrovich, Slobodanka (Yugoslavia)
Mitschi, Claude (France)
Miura, Robert M. (USA)
Mo, Zaishu (China)
Moazzami, Dara (Iran)
Model, Boris (Israel)
Moehring, Konrad W. (Germany)
Moghaddamfar, Ali Reza (Iran)
Mohammadi Hassanabadi, Aliakbar (Iran)
Mohsen, Taghavi (Iran)
Mohssine, Alif (Italy)
Mokhov, Oleg I. (Russia)
Mol, Rogério S. (Brazil)
Moldavskaya, Elina (Ukraine)
Mollov, Todor (Bulgaria)
Montans, Fernando J. (Uruguay)
Moon, Myoung Ho (Korea)
Moore, Charles N. (USA)
Moore, John D. (USA)
Moran, Gadi (Israel)
Moreira, Carlos Gustavo (Brazil)
Mori, Seiki (Japan)
Mori, Shigefumi (Japan)
Morifuji, Takayuki (Japan)
Morikuni, Goto (Japan)
Morimoto, Mitsuo (Japan)
Moriya, Katsuhiro (Japan)
Moriyama, Tetsuhiro (Japan)
Mossige, Svein (Norway)
Mostovoy, Jacob (Mexico)
Mourrain, Bernard (France)
Mousavi, Hamid (Iran)

Movshovitz-Hadar, Nitsa (Israel)
Mphako, Eunice G. (Malawi)
Mphako, Eunice Gogo (Malawi)
Mshimba, Ali Seif (Tanzania)
Mukai, Shigeru (Japan)
Mukhamedov, Farruh (Uzbekistan)
Mukherjee, Goutam (India)
Mukhtarova, Olga Vasilyevna (Azerbaijan)
Muktibodh, Arun (India)
Mumford, David B. (USA)
Munembe, Joao S. P. (Mozambique)
Muranov, Yurij Vladimirovich (Belarus)
Murasugi, Kunio (Canada)
Mustata, Mircea I. (United Kingdom)
Myung, Hyo Chul (Korea)
N, Raimo (Finland)
Nachtergaele, Bruno L. (USA)
Nagaoka, Kazuaki (Japan)
Nai, Bing (China)
Nakai, Eiichi (Japan)
Nakajima, Hiraku (Japan)
Nakajima, Toru (Japan)
Nakamura, Masaaki (Japan)
Nakamura, Yoshio (Japan)
Nakane, Michiyo (Japan)
Nakane, Shizuo (Japan)
Nam, Hoyoung (Korea)
Nam, Ki-Bong (USA)
Namdari, Mehrdad (Iran)
Namnak, Chaiwat (Thailand)
Nandakumar, Nagaiah (USA)
Napalkov, Valentin Vasilievich (Russia)
Narayanaswami, Pallasena P. (Canada)
Narici, Lawrence R. (USA)
Nash Jr., John F. (USA)
Nashed, Zuhair M. (USA)
Nassif, Coltoussoub (Canada)
Natale, Sonia L. (France)
Natarajan, Raja (India)
Nathanson, Melvyn B. (USA)
Natroshvili, David (Georgia)
Naumov, Anatoly (Russia)
Naveira, Antonio M. (Spain)
Nazarov, Maxim L. (United Kingdom)
Neammanee, Kritsana (Thailand)
Neelakanta, Sthanumoorthy (India)
Negreiros, Caio J. C. (Brazil)
Neitzke, Andrew M. (USA)
Nekrasov, Nikita Aleksandrovich (France)
Nevanlinna, Olavi (Finland)
Ng, Chi-Keung (United Kingdom)
Ng, Tuen Wai (China)
Ng, Wai-Yin (China)
Ngo, Aiet Trung (Vietnam)
Ngo, Trung V. (Vietnam)
Nguyen, Chau Van (Vietnam)
Nguyen, Dinh Tri (Vietnam)
Nguyen, Hai Hoang (Japan)
Nguyen, Q. Thang (Vietnam)
Nguyen, Tu Cuong (Vietnam)
Nguyen, Van Minh (Vietnam)
Nguyen, Viet Dung (Vietnam)
Nguyen, Xuan Tuyen (Vietnam)
Ni, Junna (China)
Ni, Lei (USA)
Ni, Qin (China)
Ni, Yi (China)
Niamsup, Piyapong (Thailand)
Nicolae, Florin (Germany)
Nicolaides, Roy A. (USA)
Nie, Puyan (China)
Nie, Zhongyi (China)
Nieto, Isidro Baños (Mexico)
Nii, Shunsaku (Japan)
Niknam, Assadollah (Iran)
Nikolic, Aleksandar M. (Yugoslavia)
Nikonorov, Yurii G. (Russia)
Nilrat, Chaufah K. (Tunesia)
Ning, Shucheng (China)
Niu, Fengwen (China)
Niu, Liang (China)
Niu, Min (China)
Njenga, Edward Gachangi (Kenya)
Nkemzi, Boniface Belagoa (Cameroon)
Noel, Alfred (USA)
Nomura, Yasutoshi (Japan)
Nongxa, Loyiso G. (South Africa)
Norbury, Paul T. (Australia)

Nordo, Giorgio (Italy)
Norton, Douglas E. (USA)
Noumi, Masatoshi (Japan)
Novak, S. Y. (United Kingdom)
Novikov, Mikhail (Russia)
Nwabueze, Kenneth Kenechukwu (Brunei)
Oakes, Susan Margaret (United Kingdom)
Obitsu, Kunio (Japan)
Obukhovskii, Valeri V. (Russia)
Odai, Yoshitaka (Japan)
O'donovan, Donal P. (Ireland)
Oeljeklaus, Eberhard (Germany)
Ogouyandjou, Carlos (Benin)
Oguiso, Keiji (Japan)
Ohba, Kiyoshi (Japan)
Ohmiya, Mayumi (Japan)
Ohnita, Yoshihiro (Japan)
Ohno, Hiloshi (Japan)
Ohtsuka, Fumiko (Japan)
Oikonomides, Catherine (Japan)
Oinarov, Ryskul (Kazakhstan)
Oka, Hiroe (Japan)
Okada, Tatsuya (Japan)
Okamoto, Yoshio (Switzerland)
Okayasu, Takateru (Japan)
Okazaki, Ryotaro (Japan)
Okuyama, Yusuke (Japan)
Oladipo, Mike Dare (Nigeria)
Olenko, Andriy Ya. (Ukraine)
Ongarbaev, Ernar Saginbekovitch (Kazakhstan)
Orlov, Dmitry Olegovich (Russia)
Orr, Kent E. (USA)
O'shea, Donal B. (USA)
Osilike, M.O. (Nigeria)
Otto, Felix (Germany)
Ou, Huikang (China)
Ouerdiane, Habib (Tunesia)
Ouyang, Caiheng (China)
Ouyang, Chongzhen (China)
Ouyang, Geng (China)
Overton, Christopher W. (USA)
Ozen, Fusun (Turkey)
Ozluk, Ali E. (USA)
Oztop, Serap (Turkey)
Paffenholz, Andreas (Germany)
Palis, Jacob (Brazil)
Palmer, William D. (Australia)
Pan, Jianfang (China)
Pan, Jian-Yu (China)
Pan, Jianzhong (China)
Pan, Jie (China)
Pan, Liping (China)
Pan, Mingyong (China)
Pan, Ping-Qi (China)
Pan, Ronghua (USA)
Pan, Shenglang (China)
Pan, Shuming (China)
Pan, Yanglian (China)
Pan, Yifei (USA)
Pan, Yong-Liang (China)
Pan, Yu (China)
Panahov, Ettbar (Turkey)
Panazzolo, Daniel (Brazil)
Pandharipande, Rahul Vijay (USA)
Pang, Shanqi (China)
Pantaragphong, Praiboon (Thailand)
Papanikolas, Matthew (USA)
Park, Donghoon (Korea)
Park, Hye. J. (Korea)
Park, Jeanam (Korea)
Park, Jongil (Korea)
Park, Kyewon K. (Korea)
Parmenter, Michael M. (Canada)
Paseman, Gerhard (USA)
Passare, Mikael (Sweden)
Payrovi, Shiroyeh (Iran)
Pechersky, Eugene (Russia)
Pedas, Arvet (Estonia)
Pei, Ruyi (China)
Pei, Zhan (USA)
Pekonen, Osmo E. T. (Finland)
Peltonen, Kirsi (Finland)
Pena, Juan M. (Spain)
Peng, Daheng (China)
Peng, Jianping (China)
Peng, Junhuan (China)
Peng, Liangang (China)

Peng, Lizhong (China)
Peng, Shuangjie (China)
Peng, Tsu-Ann (Singapore)
Peng, Wenhua (China)
Peng, Zhigang (China)
Pensupha, Luddawan (Thailand)
Peres, Yuval (USA)
Pereverzev, Sergei (Ukraine)
Perez-Esteva, Salvador (Mexico)
Perisic, Dusanka M. (Yugoslavia)
Perkins, Sarah B. (United Kingdom)
Perrine, Serge (France)
Perssan, Ulf A. (Sweden)
Petean, Jimmy (Mexico)
Peters, Martin H. (Germany)
Petersen, Sebastian J. (Germany)
Petrunin, Anton (USA)
Petrusel, Adrian (Romania)
Petrushko, Igor (Russia)
Petzold, Marko (Germany)
Peyriere, Jacques (France)
Pfeifle, Julian (Germany)
Pfister, Gerhard (Germany)
Pham, Anh Minh (Vietnam)
Pham, Anh Ngoc (Hungary)
Pham, Ky Anh (Vietnam)
Pham, The Long (Vietnam)
Pham-Gia, Thu (Canada)
Phatarfod, Ravi M. (Australia)
Philippin, Gerard A. (Canada)
Phillips, N. Christopher (USA)
Piao, Yongjie (China)
Piao, Zhihui (China)
Piatetski-Shapiro, Ilya I. (USA)
Piccione, Paolo (Italy)
Picozzi, Stefano (Switzerland)
Piene, Ragni (Norway)
Pilgrim, Kevin M. (USA)
Pilipovic, Stevan (Yugoslavia)
Pink, Richard (Switzerland)
Pinto, Alberto Adrego (Portugal)
Pisier, Gilles (France)
Pisztora, Agoston (USA)
Pizana, Romulo (Philippines)
Poetzelberger, Klaus (Austria)
Polotovsky, Grigory M. (Russia)
Pommersheim, James E. (USA)
Poovey, Mary (USA)
Popa, Eugen I. (Romania)
Popovici, Adriana (Romania)
Popovici, Dan (Romania)
Porteous, Hugh L. (United Kingdom)
Porter, R. Michael (Mexico)
Portillo-Fernández, José-Ra (Spain)
Portnoy, Arturo (USA)
Posnikov, Mikhail (Russia)
Postan, Mikhail Ya. (Ukraine)
Potapov, Vadim D. (Russia)
Pourkazemi, Mohammad Hossein (Iran)
Prabhakaran, D. J. (India)
Praeger, Cheryl E. (Australia)
Prajapat, Jyotshana (India)
Prestini, Elena (Italy)
Prieto, Carlos (Mexico)
Promislow, David (Canada)
Prommi, Prathum (Thailand)
Proskuryakov, Mikhail N. (Russia)
Ptak, Marek (Poland)
Pu, Fei (China)
Pujals, Enrique Ramiro (Brazil)
Pupo, Mauro Garcia (Cuba)
Putilina, Anna Vladimirovna (Russia)
Qi, Dongxu (China)
Qi, He (China)
Qi, Pengfei (China)
Qi, Qiulan (China)
Qi, Xu (China)
Qi, Yi (China)
Qi, Yuanwei (USA)
Qian, Meihua (China)
Qian, Neng-Sheng (China)
Qian, Xiaosong (China)
Qiao, Jianyong (China)
Qiao, Youfu (China)
Qin, Hourong (China)
Qin, Mengzhao (China)
Qin, Yun (China)
Qiu, Chunhui (China)
Qiu, Derong (China)
Qiu, Guoyong (China)

Qiu, Ruifeng (China)
Qiu, Shuxi (China)
Qiu, Tianzhen (China)
Qiu, Weisheng (China)
Qu, Anjing (China)
Qu, Changzheng (China)
Quarteroni, Alfio M. (Switzerland)
Queiroz, Marcelo G. (Brazil)
Rabia, Sherif I. (Egypt)
Rabinovitch, Avinoam (Israel)
Radulescu, Florin (USA)
Radyna, Mikalai Ya. (Belarus)
Raghunathan, Madabusi Santanam (India)
Raghunathan, Ravi (India)
Raghuram, Anantharam (India)
Rajabov, Nusrat (Tajikistan)
Rajabova, Lutfia (Tajikistan)
Rakic, Zoran (Yugoslavia)
Raman, Preeti (India)
Ramana, D. S. (India)
Rambau, Jörg (Germany)
Rannacher, Rolf (Germany)
Rao, Geetha S. (India)
Rao, Wengxing (China)
Rappoport, Juri Moiseevich (Russia)
Raslan, Kamal Raslan Mohamed (Egypt)
Rassias, Themistocles Michael (Greece)
Rasulova, Mukhayo (Uzbekistan)
Ratanapun, Suporn (Thailand)
Rathinasamy, Sakthivel (Korea)
Rattanametawee, Witchaya (Thailand)
Rattanaphet, Arisa (Thailand)
Raugel, Genevieve (France)
Rawdon, Eric J. (USA)
Raz, Ran (Israel)
Razani, Abdolrahman (Iran)
Rebiai, Salah-Eddine (Algeria)
Recke, Lutz (Germany)
Reed, Bruce Alan (Canada)
Rees, Elmer G. (United Kingdom)
Reid, Miles (United Kingdom)
Reinecke, Carolus J. (South Africa)
Ren, Bin (China)
Ren, Guanshen (USA)
Ren, Haizhen (China)
Ren, Hongshan (China)
Ren, Nanheng (China)
Ren, Xinxi (China)
Ren, Zhihua (China)
Renteria-Marquez, Carlos (Mexico)
Repovs, Dusan (Slovenia)
Reshetova, Galina (Russia)
Reutenauer, Arthur (France)
Rezaei Aliabad, Ali (Iran)
Rhodes, John A. (USA)
Ri, Myong-Hwan (DPR of Korea)
Riahi, Hasna (Tunesia)
Ricciardi, Tonia (Italy)
Richards, Franklin B. (USA)
Rickman, Seppo U. (Finland)
Rieger, Marc O. (USA)
Ringel, Claus Michael (Germany)
Riordan, Oliver M. (United Kingdom)
Rivasseau, Vincent (France)
Riviere, Tristan (Switzerland)
Riznyk, Volodymyr V. (Ukraine)
Roan, Shi-Shyr (Taiwan)
Robbiano, Lorenzo (Italy)
Roberts, Justin D. (USA)
Robinson, Bencion (Israel)
Robinson, Derek W. (Australia)
Roch, Steffen (Germany)
Rockner, Michael (Germany)
Rodkina, Alexandra (Russia)
Roe, John (USA)
Roiter, Andriy Volodymyrovych (Ukraine)
Roitman, Moshe (Israel)
Romberg, Thomas A. (USA)
Romero, Susana A. (Venezuela)
Rong, Xiaochun (USA)
Roox, Le (France)
Roquette, Peter J. (Germany)
Rordam, Mikael (Denmark)
Rosas, Mercedes H. (Venezuela)
Rossman, Wayne F. (Japan)
Rost, Markus (USA)
Rothblum, Uriel G. (Israel)

Rouhani, Behzad Djafari (Iran)
Rourke, Colin P. (United Kingdom)
Roushon, Sayed Khaled (India)
Rousseau, Christiane (Canada)
Rowley, Chris A. (United Kingdom)
Rozikov, Utkir (Uzbekistan)
Ruan, Jishou (China)
Ruan, Yongbin (USA)
Rubin, Karl (USA)
Rubinstein, Joachim H. (Australia)
Rubinstein, Zalman (Israel)
Rudolph, Daniel J. (USA)
Ruf, Bernhard (Italy)
Rui, Hebing (China)
Rukavina, Sanja (Croatia)
Ruppert, Wolfgang A. F. (Austria)
Rushing, Shelley R. (USA)
Ruzhansky, Michael (United Kingdom)
Ryashko, Lev (Russia)
Rybicki, Tomasz (Poland)
Rzedowski-Calder-n. Martha (Mexico)
Sachdev, Purushottam L. (India)
Sadallah, Boubaker (Algeria)
Sadik, Nazim (Turkey)
Sadov, Sergey (Russia)
Sadullaev, Azimbay (Uzbekistan)
Sadullaeva, Shahlo (Uzbekistan)
Sadyrbaev, Felix (Latvia)
Saez-Schwedt, Andres (Spain)
Safuanov, Ildar S. (Russia)
Sagatov, Miraziz (Uzbekistan)
Sahoo, Pravati (India)
Sahu, Dayaram (India)
Sakakibara, Nobuhisa (Japan)
Sakobiga, Salomon (China)
Sakthivel, Rathinasamy (Egypt)
Sal Moslehian, Mohammad (Iran)
Salberger, Per (Sweden)
Saleri, Fausto (Italy)
Salinger, David (United Kingdom)
Saltykov, Evgueni Grigoryevich (Russia)
Samian, Abdul Latif (Malaysia)
Samovol, Vladimir (Russia)
Samper, Carmen (Colombia)
Sanchez, Hector F. (Mexico)
Sanderson, Brian J. (United Kingdom)
Sango, Mamadou (South Africa)
Saninta, Tipaval (Thailand)
Sano, Shigeru (Japan)
Santos, David A. (USA)
Sanz Sole, Marta (Spain)
Sargolzaei, Parviz (Iran)
Sato, Hiroki (Japan)
Sato, Ken-Iti (Japan)
Sato, Kenzi (Japan)
Sato, Yumiko (Japan)
Satoh, Junya (Japan)
Satravaha, Pornchai (Thailand)
Savas, Ekrem (Turkey)
Savin, Anton Yu. (Russia)
Savitt, David Lawrence (Canada)
Sawae, Ryuichi (Japan)
Sawon, Justin (United Kingdom)
Saxena, Subhash C. (USA)
Sayed, K. Roushon (India)
Scardua, Bruno Azevedo (Brazil)
Schechtman, Vadim (France)
Scherfner, Mike (Germany)
Schmets, Jean F. H. (Belgium)
Schmidt, Dieter J.H. (Germany)
Schmitt, Bernard (France)
Schmitt, Peter (Austria)
Schmuland, Byron Allan (Canada)
Schochet, Claude (USA)
Schott, Rene (France)
Schreiber, Bertram M. (USA)
Schuster, Alexander P. (USA)
Schwab, Christoph (Switzerland)
Schwaenzl, Roland (Germany)
Schwartz, Richard E. (USA)
Sedykh, Vyacheslav Dmitrievich (Russia)
Seidel, Paul (France)
Sekita, Eitaro (Japan)
Sela, Zlil (Israel)
Senapathi, Eswara Rao (India)
Senashov, Vladimir Ivanovich (Russia)
Seo, Su Mi (Korea)
Sergeev, Armen (Russia)
Serow, Dmitry W. (Russia)

Sethian, James A. (USA)
Sha, Guoxiang (China)
Shadman, Dariush (Iran)
Shahabi Shojaei, Mohammad Ali (Iran)
Shahidi, Freydoon (USA)
Shahumyan, Harutyun (Armenia)
Shahvarani-Semnani, Ahmad (Iran)
Shamolin, Maxim V. (Russia)
Shan, Tengfeng (China)
Shan, Xiuling (China)
Shang, Weidong (China)
Shang, Yanying (China)
Shang, Yi (USA)
Shang, Zaijiu (China)
Shao, Bin (USA)
Shao, Jiayu (China)
Shao, Jinghai (China)
Shao, Qi-Man (Singapore)
Shao, Song (China)
Shao, Xiumin (China)
Shao, Zhoude (USA)
Shaposhnikova, Tatyana (Sweden)
Shapoval, Alexandr (Russia)
Sharma, Birendra Kumar (India)
Sharma, Pramod Kumar (India)
Sharma, Virendra Kumar (India)
Shawagfeh, Nabil T. (Jordan)
Shchepanyuk, Gennadiy (Ukraine)
Shchepetilov, Alexey Valerievich (Russia)
Shcherbacov, Victor Alexei (Moldova)
She, Chiqiu (China)
She, Zhikun (China)
Sheen, Dongwoo (Korea)
Shen, Hao (China)
Shen, Jianhua (China)
Shen, Jisen (China)
Shen, Junhao (China)
Shen, Ke Duan (China)
Shen, Lihua (China)
Shen, Longjun (China)
Shen, Peiping (China)
Shen, Samuel (Canada)
Shen, Weihua (China)
Shen, Weixiao (United Kingdom)
Shen, Wenxuan (China)
Shen, Yibing (China)
Shen, Yidan (China)
Shen, Yiqing (China)
Shen, Zifei (China)
Sheng, Pingxing (China)
Sheng, Qirong (China)
Sheng, Wancheng (China)
Sheng, Weimin (China)
Shepelsky, Dmitry (Ukraine)
Sherry, David (USA)
Sheth, Dilip N. (India)
Shi, Bao (China)
Shi, Dinghua (China)
Shi, Fu-Gui (China)
Shi, He (China)
Shi, Hongting (China)
Shi, Jiao-Min (China)
Shi, Jimin (China)
Shi, Kaida (China)
Shi, Lingsheng (United Kingdom)
Shi, Ming (China)
Shi, Peilin (China)
Shi, Weiping (USA)
Shi, Weixue (China)
Shi, Wujie (China)
Shi, Xianliang (China)
Shi, Yingjie (China)
Shi, Yiqian (China)
Shi, Yufeng (China)
Shi, Yuguang (China)
Shi, Yuying (China)
Shi, Zhongci (China)
Shi, Zhongrui (China)
Shibano, Hiroki (Japan)
Shilnikov, Leonid Pavlovich (Russia)
Shimada, Nobuo (Japan)
Shin, Dong-Kwan (Korea)
Shindin, Sergey Konstantinovich (Russia)
Shintani, T. (Japan)
Shiohama, Katsuhiro (Japan)
Shiraiwa, Kenichi (Japan)
Shishikura, Mitsuhiro (Japan)
Shiue, Peter (USA)

Shliahov, Vladislav Viktorovich (Ukraine)
Shmatkov, Ruslan Nikolaevich (Russia)
Sho, Moon-Kyung (Korea)
Shpenkov, George P. (Poland)
Shu, Bin (China)
Shu, Chi-Wang (USA)
Shu, Lin (China)
Shu, Lisheng (China)
Shuai, Zhisheng (China)
Shubin, Mikhail A. (USA)
Shunkov, Vladimir Petrovich (Russia)
Shutyaev, Victor (Russia)
Si, Jianguo (China)
Sibanda, Precious (Zimbabwe)
Siddiqi, Abul Hasan (Saudi Arabia)
Siebenmann, Laurent C. (France)
Sigmund, Karl H. (Austria)
Signoret, Carlos J. (Mexico)
Sim, Chol (Korea)
Simons, Gordon E. (Canada)
Singh, Anand Prakash (India)
Singh, Mansa C. (Canada)
Sinha, Kalyan B. (India)
Sintamarian, Alina (Romania)
Siren, Daoreji (China)
Sirvent, Victor F. (Venezuela)
Siu, Yum-Tong (USA)
Sizikov, V. S. (Russia)
Skill, Thomas (Germany)
Skoda, Zoran (USA)
Skopenkov, Arkadi (Russia)
Skrypnik, Ihor Vladimirovich (Ukraine)
Skubachevskii, Alexander L. (Russia)
Slijepcevic, Sinisa (Czech)
Sloan, Ian H. (Australia)
Smagulov, Shaltay (Kazakhstan)
Smailov, Yesmukhanbet S. (Kazakhstan)
Smarandache, Florentin (USA)
Smillie, John (USA)
Smith, Gregory G. (USA)
Smith, Pamela F. (USA)
Smith, Stuart P. (USA)
Smoluk, Antoni (Poland)
Snoussi, Jawad (Mexico)
Soares, M. G. (Brazil)
Sobolev, Vladimir (Russia)
Soekirno, Ichary (Indonesia)
Sofi, Mohd. Amin (India)
Soheili, Ali Reza (Iran)
Solarin, Adewale (Nigeria)
Solomyak, Boris (USA)
Sombra, Martin A. (France)
Son, Gyoyong (Korea)
Son, Lehung (Vietnam)
Son, Nguyen Khoa (Vietnam)
Song, Baorui (China)
Song, Binheng (China)
Song, Guoqiang (China)
Song, Hao (China)
Song, Lixin (China)
Song, Meimei (China)
Song, Min (China)
Song, Ming (China)
Song, Ruikun (China)
Song, Ruowei (China)
Song, Shichang (China)
Song, Shiji (China)
Song, Shukui (China)
Song, Wen (China)
Song, Yongjin (Korea)
Song, Yongzhong (China)
Song, Yu (China)
Song, Zhenming (China)
Sos, Vera (Hungary)
Soulé, Christophe J. (France)
Spatzier, Ralf J. (USA)
Spector, Lawrence B. (USA)
Speed, Terence Paul (USA)
Spielman, Daniel Alan (USA)
Spring, David H. (Canada)
Srivastava, Tariq (Canada)
St. Mary, Donald F. (USA)
Stafford, J. Toby (USA)
Stamatovic, Biljana (Yugoslavia)
Stanisic, Predrag (Yugoslavia)
Steger, Tim Joshua (Italy)
Stein, Greg (USA)
Steinberger, Mark (USA)
Steinhorn, Charles I. (USA)

Stempien, Zdzislaw (Poland)
Stenlund, Mikko S. (Finland)
Sternheimer, Daniel H. (France)
Stewart, Cameron L. (Canada)
Storm, Peter A. (USA)
Storozhev, Valery I. (Ukraine)
Strano, Rosario (Italy)
Strauch, Matthias (Germany)
Stray, Arne (Norway)
Stroth, Gernot (Germany)
Strunkov, S. P. (Russia)
Su, Buqing (China)
Su, Hongling (China)
Su, Huaming (China)
Su, Jiabao (China)
Su, Jianbing (China)
Su, Li Jie (China)
Su, Ning (China)
Su, Xiaole (China)
Su, Xiaoquan (China)
Su, Yalatu (China)
Su, Yang (China)
Su, Yucai (China)
Su, Zhong Gen (China)
Sudan, Madhu (USA)
Sudo, Masaki (Japan)
Suffridge, Ted J. (USA)
Sugawara, Tamio (Japan)
Suh, Dong Youp (Korea)
Sukhotin, Alexander Mikhailovich (Russia)
Sukla, Indu Lata (India)
Sullivan, John M. (USA)
Sun, Chunyou (China)
Sun, Cuifang (China)
Sun, Daode (China)
Sun, Fangyu (China)
Sun, Guozheng (China)
Sun, He (China)
Sun, Hejun (China)
Sun, Huafei (China)
Sun, Hui (China)
Sun, Jiachang (China)
Sun, Jianhua (China)
Sun, Jianhua (Nanjing) (China)
Sun, Jianqiang (China)
Sun, Jie (China)
Sun, Kai (China)
Sun, Lijuan (China)
Sun, Naizhe (China)
Sun, Nigang (China)
Sun, Peng (China)
Sun, Ping (China)
Sun, Qi (China)
Sun, Ruitao (China)
Sun, Shanli (China)
Sun, Shanzhong (China)
Sun, Shunhua (China)
Sun, Wenchang (China)
Sun, Wenjun (China)
Sun, Wenxiang (China)
Sun, Xiaotao (China)
Sun, Xingping (USA)
Sun, Yajuan (China)
Sun, Yanfeng (China)
Sun, Yeneng (Singapore)
Sun, Yidong (China)
Sun, Yongzhong (China)
Sun, Yongzhong (China)
Sun, Young John (China)
Sun, Yun (China)
Sun, Zhaocai (China)
Sun, Zhi Wei (China)
Sun, Zhihong (China)
Sun, Zhiren (China)
Sun, Zongming (China)
Sunada, Toshikazu (Japan)
Sunder, Viakalathur S. (India)
Sunley, Judith S. (USA)
Svrtan, Dragutin (Croatia)
Swanson, Irena (USA)
Sy, Polly Wee (Philippines)
Szabo, Zoltan I. (Oman)
Szalay, Laszlo (Germany)
Szemeredi, Endre (USA)
Szymanski, Waclaw (USA)
Tadmor, Eitan (USA)
Taghavi, Mohsen (Iran)
Taguchi, Yuichiro (Japan)
Tahri, El Hassan (Morocco)

Tai, Yongming (China)
Taimanov, Iskander (Russia)
Takayama, Manabu (Japan)
Takenouchi, Osamu (Japan)
Takeo, Fukiko (Japan)
Takesaki, Masamichi (USA)
Tamamura, Akie (Japan)
Tamarkin, Dmitry E. (USA)
Tamarkin, Dmitry E. (USA)
Tamaru, Hiroshi (Japan)
Tamm De Araujo Moreira, Carlos Gustavo (Brazil)
Tamrazov, Promarz (Ukraine)
Tan, Changmei (China)
Tan, Choon Ee Roger (Singapore)
Tan, Jieqing (China)
Tan, Liang (China)
Tan, Tianrong (China)
Tan, Xiao-Jiang (China)
Tanahashi, Kotaro (Japan)
Tanasi, Corrado (Italy)
Tandon, Rajat (India)
Tang, Chunlei (China)
Tang, Francis C.Y. (USA)
Tang, Gaohua (China)
Tang, Gong-You (China)
Tang, Guoping (China)
Tang, Hongmin (China)
Tang, Jiangang (China)
Tang, Jianshan (China)
Tang, Jin (China)
Tang, Lin (China)
Tang, Min (China)
Tang, Moxun (USA)
Tang, Qinggan (China)
Tang, Sanping (China)
Tang, Sanyi (China)
Tang, Shanjian (China)
Tang, Sheng-qiang (China)
Tang, Tao (China)
Tang, Xiang (USA)
Tang, Xiaomin (China)
Tang, Yanbin (China)
Tang, Yifa (China)
Tang, Yun (China)
Tang, Zhengquan (China)
Tang, Zhongwei (China)
Tangjitwatanakul, Siriwan (Thailand)
Taniguchi, Masaharu (Japan)
Tao, Dongqing (China)
Tao, Xiangxing (China)
Tao, Yongqian (China)
Tarasov, Vitaly (Russia)
Tataru, Daniel I. (USA)
Tavakoli, Javad (Canada)
Taylor, Martin J. (United Kingdom)
Taylor, Richard Lawrence (USA)
Taymanov, Iskander (Russia)
Tchernykh, Elena (Russia)
Tcheverda, Vladimir (Russia)
Tee, Kah Ling (Malaysia)
Teicher, Mina (Israel)
Teichmann, J. (Austria)
Teichner, Peter (USA)
Tena, Juan (Spain)
Terai, Nobuhiro (Japan)
Termwuttipong, Imchit (Thailand)
Thalmaier, Anton (Germany)
Thangadurai, Ravindranathan (India)
Thao, Vo Dang (Vietnam)
Thas, Joseph Adolphe (Belgium)
Thera, Michel (France)
Thiele, Christoph (USA)
Thomas, Bouetou Bouetou (Cameroon)
Thomas, Charles B. (United Kingdom)
Thompson, Anthony (Canada)
Thompson, Robert J. (USA)
Thomsen, Momme Johs (Germany)
Thorbjornsen, Steen E. (Denmark)
Tian, Dong (China)
Tian, Fanji (China)
Tian, Gang (USA)
Tian, Haiyan (China)
Tian, Junzhong (China)
Tian, Tingyan (China)
Tian, Yimin (China)
Tian, Yubin (China)
Tian, Zhenfu (China)
Tian, Zheng (China)
Tian, Zihong (China)

Tiep, Pham Huu (USA)
Tikhonov, Sergey V. (Belarus)
Tillmann, Ulrike Luise (United Kingdom)
Timoney, Richard M. (Ireland)
Tirumalasetty, Amaranath (India)
Toda, Magdalena D. (USA)
Tojeiro, Ruy (Brazil)
Tokizawa, Masamichi (Japan)
Tomatsu, Reiji (Japan)
Tomiyama, Jun J.T. (Japan)
Tonegawa, Yoshihiro (Japan)
Tonev, Thomas (Toma) (USA)
Tong, Hui (China)
Tong, Zengxiang (USA)
Ton-That, Tuong (USA)
Torisu, Ichiro (Japan)
Torre, Anna (Italy)
Torres, Pedro J. (Spain)
Torres, Rodolfo H. (USA)
Totaro, Burt (United Kingdom)
Touraev, Vladimir G. (France)
Tovar-Sanchez, Luis Manuel (Mexico)
Tracy, Craig Arnold (USA)
Tranah, David A. (United Kingdom)
Traves, William N. (USA)
Traynor, Tim (Canada)
Trenogin, Vladilen A. (Russia)
Treschev, Dmitry (Russia)
Trivisa, Konstantina (USA)
Troitsky, Evgenij V. (Russia)
Trudinger, Neil S. (Australia)
Tsai, Chung-Ju (New Zealand)
Tseng, Hsian-Hua (USA)
Tserennadmid, Batkhuu (Mongolia)
Tsoodol, Dolgorsuren (Mongolia)
Tsuboi, Takashi (Japan)
Tsushima, Ryuji (Japan)
Tuneski, Nikola (Macedonia)
Turski, Jacek (USA)
Turunen, Ville P. (Finland)
Tverberg, Helge (Norway)
Tyulyukin, Vladimir A. (Russia)
Uchiyama, Mitsuru (Japan)
Udriste, Constantin (Romania)
Ueno, Kenji (Japan)
Ueno, Yoshiaki (Japan)
Uglanov, Alexey V. (Russia)
Ulecia, Teresa (Spain)
Ullmo, Emmanuel B. (France)
Um, Ko-Woon (Korea)
Upmeier, Harald (Germany)
Ushijima, Akira (Japan)
Usoltsev, Lev Pavlovich (Russia)
Uuye, Otgonbayar (Japan)
Vaamonde, Antonio (Spain)
Vainikko, Gennadi (Finland)
Vakil, Ravi (USA)
Vakily, Ghodsieh (Iran)
Valentina, Golubeva (Russia)
Valtr, Pavel (Czech)
Vâmos, Peter (United Kingdom)
Van Der Kallen, Wilberd L. (Netherlands)
Van Der Poorten, Alfred J. (Australia)
Vanegas, Carmen J. (Venezuela)
Varsaie, Saad (Iran)
Vasco, Carlos E. (Colombia)
Vasiliev, Alexander (Chile)
Vasiliev, Oleg V. (Russia)
Vasilieva, Olga (Colombia)
Vasin, Dmitrii V. (Russia)
Vasquez-Martinez, Claudio-Rafael (Mexico)
Vdovin, Evgeni P. (Italy)
Vella, Antoine (Canada)
Vélu, Jacques (France)
Venkataramana, Tyakal N. (India)
Verbovskiy, Viktor (Kazakhstan)
Vergne, Michele (France)
Verma, Jugal K. (India)
Vernaeve, Hans (Belgium)
Vershinin, Vladimir V. (France)
Viana, Marcelo (Brazil)
Viano, Juan (Spain)
Victoria, Redina M. (Philippines)
Vidal Rodeiro, Carmen L. (Spain)
Vidyasagar, Mathukumalli (India)
Viehweg, Eckart (Germany)
Vigneras, Marie (France)

Villari, Gabriele (Italy)
Vinogradov, Oleg Pavlovich (Russia)
Viola, Carlo (Italy)
Viwatwongkasem, Chukiat (Thailand)
Vodopianov, Serguei K. (Russia)
Voevodsky, Vladimir (USA)
Vogtmann, Karen (USA)
Voigt, Thomas (Germany)
Volevich, Leonid R. (Russia)
Von Mouche, Pierre (Netherlands)
Vukovic, Mirjana (Bosnia and Herzegovina)
Wada, Tomoyuki (Japan)
Waencharoen, Sribudh (Thailand)
Wahl, Elizabeth (Uzbekistan)
Wahl, Jonathan M. (USA)
Wahyuni, Sri (Indonesia)
Wakabayashi, Isao (Japan)
Walden, Byron L. (USA)
Waldschmidt, Michel (France)
Wales, David B. (USA)
Wallin, Hans E. (Sweden)
Walthoe, Jonathan M. (United Kingdom)
Wan, Daqing (USA)
Wan, Liangxia (China)
Wan, Zhexian (Canada)
Wang, An (China)
Wang, Baoshan (China)
Wang, Bin (China)
Wang, Binglin (China)
Wang, Caitlin Y. (USA)
Wang, Chang You (USA)
Wang, Changce (China)
Wang, Changping (China)
Wang, Chao (China)
Wang, Chiew Peng (Malaysia)
Wang, Chunjie (China)
Wang, Chunrun (China)
Wang, Delin (China)
Wang, Desheng (China)
Wang, Dingkang (China)
Wang, Dong (China)
Wang, Dong Qian (New Zealand)
Wang, Dongming (China)
Wang, Duo (China)
Wang, Fangting (China)
Wang, Feng (China)
Wang, Fengyu (China)
Wang, Fuzheng (China)
Wang, Gang (China)
Wang, Gang (USA)
Wang, Geping (China)
Wang, Gongbao (China)
Wang, Guangming (China)
Wang, Guilan (China)
Wang, Guolian (China)
Wang, Haihui (China)
Wang, Haiming (China)
Wang, Han (China)
Wang, Hanxing (China)
Wang, Haohao (USA)
Wang, Heping (China)
Wang, Hong (Canada)
Wang, Hong (China)
Wang, Hong-Ji (China)
Wang, Hongxia (China)
Wang, Hongyu (China)
Wang, Huayang (China)
Wang, Hui (China)
Wang, Huijuan (China)
Wang, Jian (China)
Wang, Jianfang (China)
Wang, Jianjun (China)
Wang, Jianpan (China)
Wang, Jiaping (USA)
Wang, Jiayin (China)
Wang, Jie (China)
Wang, Jie (BTI) (China)
Wang, Jing (China)
Wang, Jinghua (China)
Wang, Jingtao (China)
Wang, Jun (China)
Wang, Jun (Shanghai) (China)
Wang, Junping (China)
Wang, Junqing (China)
Wang, Juping (China)
Wang, Kai (China)
Wang, Ke (China)

Wang, Kunpeng (China)
Wang, Kunyang (China)
Wang, Lanyu (China)
Wang, Li (China)
Wang, Lie (China)
Wang, Lieheng (China)
Wang, Liming (China)
Wang, Linshu (China)
Wang, Liping (China)
Wang, Liqun (Canada)
Wang, Liyun (China)
Wang, Lizhong (China)
Wang, Long (China)
Wang, Long L. (USA)
Wang, Maofa (China)
Wang, Minghe (China)
Wang, Ming-Sheng (China)
Wang, Ming-Yan (China)
Wang, Naishi (China)
Wang, Naiyan (China)
Wang, Pengtao (China)
Wang, Qihua (China)
Wang, Qin (China)
Wang, Qinggang (China)
Wang, Qingwen (China)
Wang, Qingzheng (China)
Wang, Quan Fang (China)
Wang, Ruiqi (China)
Wang, Ruji (China)
Wang, Ruliang (China)
Wang, Sannuan (China)
Wang, Sheng (China)
Wang, Shi Kun (China)
Wang, Shicheng (China)
Wang, Shikun (China)
Wang, Shin-Hwa (Taiwan)
Wang, Shiwei (China)
Wang, Shugui (China)
Wang, Shuqin (China)
Wang, Silei (China)
Wang, Suyun (China)
Wang, Tao (China)
Wang, Tianze (China)
Wang, Tongchao (China)
Wang, Tonggen (China)
Wang, Wei (China)
Wang, Wei (Nanjing) (China)
Wang, Weicheng (Taiwan)
Wang, Weiqiang (USA)
Wang, Wendi (China)
Wang, Wensheng (China)
Wang, Xiaodi (USA)
Wang, Xiaofeng (China)
Wang, Xiaofeng (Sichuan) (China)
Wang, Xiaomeng (China)
Wang, Xiaomin (China)
Wang, Xiaoming (USA)
Wang, Xiaoping (China)
Wang, Xiaoqian (China)
Wang, Xiaoyu (China)
Wang, Xin (China)
Wang, Xinghua (China)
Wang, Xinping (China)
Wang, Xu-Jia (Australia)
Wang, Ya-Guang (China)
Wang, Yan (China)
Wang, Yanbin (China)
Wang, Yan-Fei (China)
Wang, Yannan (China)
Wang, Yanxin (China)
Wang, Yanying (China)
Wang, Yaodong (China)
Wang, Yejuan (China)
Wang, Yi (China)
Wang, Ying (China)
Wang, Yong (China)
Wang, Yongge (China)
Wang, Yonghui (China)
Wang, Youde (China)
Wang, Youning (China)
Wang, Yuan (China)
Wang, Yuandi (China)
Wang, Yuanhua (China)
Wang, Yuefei (China)
Wang, Yusheng (China)
Wang, Yushun (China)
Wang, Yuwen (China)
Wang, Zhaochong (China)
Wang, Zhaojun (China)
Wang, Zhengdong (China)

Wang, Zhenhui (China)
Wang, Zhigang (China)
Wang, Zhiguo (China)
Wang, Zhixi (China)
Wang, Zhong (China)
Wang, Zhongqiang (China)
Wang, Zikun (China)
Wang , Zhiqiang (USA)
Wanka, Gert (Germany)
Wanner, Gerhard (Switzerland)
Wassmer, Arnold J. (Germany)
Watanabe, Toshihiro (Japan)
Watson, Richard Oliver (Ireland)
Watt, Stephen M. (Canada)
Webb, Jeffrey R. (United Kingdom)
Weder, Ricardo A. (Mexico)
Wegner, Bernd (Germany)
Wei, Baoshe (China)
Wei, Cuiping (China)
Wei, Fajin (China)
Wei, Fengying (China)
Wei, Guofang (USA)
Wei, Jianying (China)
Wei, Jinhe (China)
Wei, Li (China)
Wei, Liping (China)
Wei, Shihshu Walter (USA)
Wei, Shuyun (China)
Wei, Wenbin (China)
Wei, Wenzhang (China)
Wei, Wu (China)
Wei, Xianhua (China)
Wei, Yimin (China)
Wei, Zhongli (China)
Weiss, Asia I. (Canada)
Weiss, Gary (USA)
Weit, Yitzhak (Israel)
Welch, Amy E. (USA)
Welch, Philip (Germany)
Wen, Bangyan (China)
Wen, Lan (China)
Wen, Songlong (China)
Wen, Xianzhang (China)
Wen, Yangyang (China)
Wen, Zhiying (China)
Wencel, Roman (Poland)
White, Brian Cabell (USA)
Wieczorek, Wojciech (USA)
Wiegand, Sylvia (USA)
Wilcox, Diane (South Africa)
Williams, Lauren K. (USA)
Wilson, Pelham M. (United Kingdom)
Winitsky De Spinadel, Vera Martha (Argentina)
Winkler, Peter (USA)
Witten, Edward (USA)
Wittmann, Christian (Germany)
Wittum, Gabriel C. (Germany)
Wojtkowski, Maciej P. (USA)
Wong, Ngai-Ching (Taiwan)
Wong, Raymond Y. (USA)
Wong, Roderick S. C. (China)
Wong, Shiu-Chun (China)
Wood, Aihua (USA)
Woodin, Hugh (USA)
Wooley, Trevor Dion (USA)
Wright, James R. (United Kingdom)
Wschebor, Mario (Uruguay)
Wu, Aiwen (China)
Wu, Baoqiang (China)
Wu, Baoyinduren (China)
Wu, Chuanxi (China)
Wu, Chuntao (USA)
Wu, Congxin (China)
Wu, Da (China)
Wu, Duzhi (China)
Wu, Faen (China)
Wu, Guangyao (China)
Wu, Guofu (China)
Wu, Guohua (New Zealand)
Wu, Haidong (USA)
Wu, Haijun (China)
Wu, Hansheng (Japan)
Wu, Huoxiong (China)
Wu, Jianhong (Canada)
Wu, Jinbiao (China)
Wu, Jinrong (China)
Wu, Ke (China)
Wu, Liming (China)
Wu, Lingyun (China)

Wu, Longqing (China)
Wu, Mengda (China)
Wu, Min (China)
Wu, Ming (China)
Wu, Quanshui (China)
Wu, Qunying (China)
Wu, Runheng (China)
Wu, Shengjian (China)
Wu, Shuhong (China)
Wu, Sijue (USA)
Wu, Siye (USA)
Wu, Songlin (China)
Wu, Taoyang (China)
Wu, Tieru (China)
Wu, Wei (China)
Wu, Weixing (China)
Wu, Wen-Tsun (China)
Wu, Wenqing (China)
Wu, Xiaoming (China)
Wu, Xinsheng (China)
Wu, Xinwen (China)
Wu, Yaokun (China)
Wu, Yaping (China)
Wu, Yueming (USA)
Wu, Yujiang (China)
Wu, Yunan (China)
Wu, Zemin (China)
Wu, Zhaoji (China)
Wu, Zhen (China)
Wu, Zhengpeng (China)
Wu, Zhihua (China)
Wu, Zhiyu (China)
Wulan, Hasi (China)
Wulfsohn, Aubrey (United Kingdom)
Xi, Nanhua (China)
Xi, Xianghua (China)
Xi, Zairong (China)
Xia, Bican (China)
Xia, Chang Yu (Brazil)
Xia, Jianming (China)
Xia, Jing (China)
Xia, Zhihong (USA)
Xian, Ming (China)
Xian, Qiuhong (USA)
Xiang, Kai-Yao (China)
Xiang, Shuhuang (China)
Xiao, Hongyi (China)
Xiao, Hongying (China)
Xiao, Jie (China)
Xiao, Liangliang (China)
Xiao, Liuqing (China)
Xiao, Mengqiu (China)
Xiao, Shutie (China)
Xiao, Tingyan (China)
Xiao, Xiaoyu (China)
Xie, Bin (China)
Xie, Dezheng (China)
Xie, Feng (China)
Xie, Ganquan (USA)
Xie, Lin (China)
Xie, Quanbo (Germany)
Xie, Shijie (China)
Xie, Shishen (USA)
Xie, Xiaoping (China)
Xie, Yingchao (China)
Xie, Yinling (China)
Xie, Zhaoru (China)
Xie, Ziqing (China)
Xin, Cunyan (China)
Xin, Xiaodong (China)
Xin, Yumei (China)
Xin, Zhouping (China)
Xing, Chaoping (Singapore)
Xing, Xiuxia (China)
Xing, Yongsheng (China)
Xiong, Chunguang (China)
Xiong, Daguo (China)
Xiong, Guohua (China)
Xu, Bin (Japan)
Xu, Caiyong (China)
Xu, Chuan-Ju (China)
Xu, Da (China)
Xu, Dachuan (China)
Xu, Deliang (China)
Xu, Delong (China)
Xu, Dinghua (China)
Xu, Fei (China)
Xu, Fuhua (China)
Xu, Guangshan (China)
Xu, Guiqiao (China)

Xu, Guoliang (China)
Xu, Hongwei (China)
Xu, Jiancheng (China)
Xu, Jianhua (China)
Xu, Jin (China)
Xu, Jinchao (USA)
Xu, Kang (China)
Xu, Kejian (China)
Xu, Luoshan (China)
Xu, Manling (China)
Xu, Meng (China)
Xu, Mingyao (China)
Xu, Minhui (China)
Xu, Pei (USA)
Xu, Pengcheng (China)
Xu, Shangjin (China)
Xu, Shengzhi (China)
Xu, Shimeng (China)
Xu, Shuzhan (China)
Xu, Tong (China)
Xu, Wei (China)
Xu, Wenhao (China)
Xu, Xianmin (China)
Xu, Xiaoshan (China)
Xu, Xingzhong (China)
Xu, Xuejun (China)
Xu, Yuan-Tong (China)
Xu, Yueheng (China)
Xu, Yuesheng (China)
Xu, Yun (China)
Xu, Yunge (China)
Xu, Zhongling (USA)
Xu, Zuoliang (China)
Xu, Zuoquan (China)
Xuan, Pei Cai (China)
Xuan, Xiaohua (China)
Xuan, Yulin (China)
Xue, Qingying (China)
Xue, Xinwei (China)
Xue, Yan (China)
Xue, Yunhua (China)
Yakovlev, Sergey (Ukraine)
Yamaguchi, Tadashi (Japan)
Yamanaka, Takeshi (Japan)
Yamanoshita, Tsuneyo (Japan)
Yamasaki, Masayuki (Japan)
Yamashita, Go (Japan)
Yamoah, David K. (Ghana)
Yan, Catherine H. (USA)
Yan, Congquan (China)
Yan, Dunyan (China)
Yan, Gui Ying (China)
Yan, Guoguang (China)
Yan, Guojun (China)
Yan, Haifeng (China)
Yan, Huahui (China)
Yan, Jia-An (China)
Yan, Jurang (China)
Yan, Kuihua (China)
Yan, Kuiying (China)
Yan, Min (China)
Yan, Ning Ning (China)
Yan, Ping (China)
Yan, Qiantai (China)
Yan, Qiping (China)
Yan, Song Y.(China)
Yan, Wei (China)
Yan, Xia (China)
Yan, Ying (China)
Yan, Yinhui (China)
Yanaba, Hiroko (Japan)
Yang, Bicheng (China)
Yang, Bizhong (China)
Yang, Changsen (China)
Yang, Chao (China)
Yang, Chunhong (China)
Yang, Cuihong (China)
Yang, Dachun (China)
Yang, Deyun (China)
Yang, Dong (China)
Yang, Fenghong (China)
Yang, Fuchun (China)
Yang, Fuzhong (China)
Yang, Guijun (China)
Yang, Guoxiao (China)
Yang, Haicheng (China)
Yang, Hailiang (China)
Yang, Haitao (China)
Yang, Han (China)
Yang, Hanchun (China)

Yang, Hongcang (China)
Yang, Hu (China)
Yang, Jae-Hyun (Korea)
Yang, Jianfu (China)
Yang, Jianping (China)
Yang, Jinghua (China)
Yang, Jun (China)
Yang, Jun (Lanzhou) (China)
Yang, Junwei (China)
Yang, Le (China)
Yang, Lei (China)
Yang, Lianzhong (China)
Yang, Libo (China)
Yang, Lijun (China)
Yang, Limin (China)
Yang, Ling E (China)
Yang, Lingling (China)
Yang, Paul C. (USA)
Yang, Qigui (China)
Yang, Qingzhi (China)
Yang, Qixiang (China)
Yang, Renzi (China)
Yang, Shang-Jun (China)
Yang, Shuo (China)
Yang, Su-Win (Taiwan)
Yang, Wanli (China)
Yang, Weiqiang (China)
Yang, Wenmao (China)
Yang, Xiangqun (China)
Yang, Xiaojing (China)
Yang, Xiaoping (China)
Yang, Xiaozhou (China)
Yang, Xin (China)
Yang, Xinping (China)
Yang, Xuan (China)
Yang, Xunnian (China)
Yang, Yang (China)
Yang, Yong (China)
Yang, Yongshou (China)
Yang, Yongzhi (USA)
Yang, Zhijie (China)
Yang, Zhiqing (USA)
Yao, Guangtong (China)
Yao, Guowu (China)
Yao, Jiangang (China)
Yao, Savage (China)
Yao, Yi (China)
Yao, Yi-Jun (China)
Yao, Ying (China)
Yao, Yiyi (China)
Yao, Yunfei (China)
Yao, Zhengan (China)
Yaroshevsky, Vassily A. (Russia)
Yaschenko, Ivan (Russia)
Yasemin, Bahar (Turkey)
Ye, Azhong (China)
Ye, Guoju (China)
Ye, Huang (China)
Ye, Peixin (China)
Ye, Qiang (Singapore)
Ye, Qixiao (China)
Ye, Shanli (China)
Ye, Xiangdong (China)
Ye, Xingde (China)
Ye, Yangbo (USA)
Ye, Yu (China)
Ye, Zhonghao (China)
Ye, Zhongxing (China)
Yeh, Tzung-Shin (China)
Yeh, Yeong-Nan (Taiwan)
Yengibaryan, Norayr B. (Armenia)
Yeung, Chi Yung (China)
Yeung, David Yiu Pik (Canada)
Yi, Lijun (China)
Yi, Zhong (China)
Yin, Hao (China)
Yin, Linsheng (China)
Yin, Lüyuan (China)
Yin, Weiping (China)
Yin, Xianjun (Germany)
Yin, Xiaobin (China)
Yin, Yanmin (China)
Yin, Yongcheng (China)
Yin, Yujuan (China)
Ying, Jiangang (China)
Ying, Longan (China)
Ylinen, Kari E. (Finland)
Yoccoz, J.C. (France)
Yokoyama, Shin (Japan)
Yoo, Ki-Jo (Korea)

Yooyuanyong, Suabsagun (Thailand)
Yoshida, Takahiko (Japan)
Yoshida, Uichi (Japan)
Yoshimoto, Akinori (Japan)
Yoshinaga, Masahiko Y. (Japan)
Yoshino, Masafumi (Japan)
Yoshino, Taro (Japan)
You, Hong (China)
You, Shihui (China)
Young, Todd Ray (USA)
Youssef, Maged Z. (Egypt)
Yu, Bo (China)
Yu, Dapeng (China)
Yu, Dehao (China)
Yu, Desheng (China)
Yu, Feng-Yuan (China)
Yu, Guoliang (USA)
Yu, Haibo (China)
Yu, Haixiao (China)
Yu, Hong (China)
Yu, Hui (China)
Yu, Jianmin (China)
Yu, Jianshe (China)
Yu, Jiayong (China)
Yu, Jiehong (China)
Yu, Jinghu (China)
Yu, Jun Hou (China)
Yu, Junyang (China)
Yu, Long (China)
Yu, Mei (China)
Yu, Pei (Canada)
Yu, Qiang (China)
Yu, Qigang (China)
Yu, Shuxiang (China)
Yu, Tao (China)
Yu, Tianran (China)
Yu, Weiwei (China)
Yu, Xintang (China)
Yu, Yinlong (China)
Yu, Yong-Guang (China)
Yu, Zenghai (China)
Yu, Zhiling (China)
Yuan, Binxian (China)
Yuan, Demei (China)
Yuan, Guangwei (China)
Yuan, Hongjun (China)
Yuan, Hui-Ping (China)
Yuan, Jianhua (China)
Yuan, Li (China)
Yuan, Mingli (China)
Yuan, Xiangdong (China)
Yuan, Xiaoping (China)
Yuan, Ya-Xiang (China)
Yuan, Yu (USA)
Yuan, Yuan (China)
Yuan, Zhijian (Finland)
Yue, Chengbo (USA)
Yue, Qin (China)
Yue, Xingye (China)
Yukich, Joseph E. (USA)
Yun, Baoqi (China)
Yun, Jin Sun (Korea)
Yun, Zhiwei (China)
Yunusi, Mahmadsyusuf Kamarzoda (Tajikistan)
Zafarani, Jafar (Iran)
Zahedi, Mohammad Mehdi (Iran)
Zaitsev, Andrei Yu. (Russia)
Zaitsev, Iouri (Brazil)
Zajac, Michal (Slovakia)
Zakharov, Valeri Konstantinovich (Russia)
Zalcman, Lawrence (Israel)
Zan, Dee (China)
Zapala, August Michal (Poland)
Zavalishchin, Dmitrii Stanislav (Russia)
Zayed, Ahmed I. (USA)
Zdravkovska, Smilka (USA)
Zeilberger, Doron (USA)
Zeitouni, Ofer (Israel)
Zelazko, Wieslaw T. (Poland)
Zelditch, Steven (USA)
Zelenyuk, Yevhen (Ukraine)
Zemanek, Jaroslav (Poland)
Zeng, Guangxing (China)
Zeng, Jiang (France)
Zeng, Jin (China)
Zeng, Jiwen (China)
Zeng, Quanping (China)
Zeng, Wenyi (China)

Zeng, Xiaoming (China)
Zeng, Yuesheng (China)
Zeng, Yunpo (China)
Zeng, Zhigang (China)
Zhan, Huirong (China)
Zhan, Shilin (China)
Zhan, Tao (China)
Zhan, Xingzhi (China)
Zhan, Zhongwei (China)
Zhang, Baoshan (China)
Zhang, Baoxue (China)
Zhang, Bin (China)
Zhang, Bingxia (China)
Zhang, Bo (United Kingdom)
Zhang, Canming (China)
Zhang, Changchun (China)
Zhang, Chao (Canada)
Zhang, Cheng (China)
Zhang, Chengbin (China)
Zhang, Chengguo (China)
Zhang, Chuang (China)
Zhang, Chuanyi (China)
Zhang, Chungou (China)
Zhang, Congjun (China)
Zhang, Dajun (China)
Zhang, Dansong (China)
Zhang, Dexue (China)
Zhang, Duanzhi (China)
Zhang, Dun-Mu (China)
Zhang, Fa-Yong (China)
Zhang, Fubao (China)
Zhang, Fuji (China)
Zhang, Fuxi (China)
Zhang, Guang (China)
Zhang, Guanglian (China)
Zhang, Guangxiang (China)
Zhang, Guangyuan (China)
Zhang, Guanquan (China)
Zhang, Guichang (China)
Zhang, Guoming (China)
Zhang, Guoyin (China)
Zhang, Hanfang (China)
Zhang, Hanqin (China)
Zhang, Heping (China)
Zhang, Hongbiao (China)
Zhang, Hongqing (China)
Zhang, Huai (China)
Zhang, Huaiyu (China)
Zhang, Huayun (China)
Zhang, Hui (China)
Zhang, Huijun (China)
Zhang, James J. (USA)
Zhang, Jiangfeng (China)
Zhang, Jianjun (China)
Zhang, Jie (China)
Zhang, Jie-Lin (China)
Zhang, Jifeng (China)
Zhang, Ji-Huan (China)
Zhang, Jin (Canada)
Zhang, Jinchuan (China)
Zhang, Jingxiao (China)
Zhang, Jingzhong (China)
Zhang, Jinxia (China)
Zhang, Jinyu (China)
Zhang, Jinzhai (China)
Zhang, Jiping (China)
Zhang, John L. (USA)
Zhang, Juliang (China)
Zhang, Jun (USA)
Zhang, Junhua (China)
Zhang, Kai-Jun (China)
Zhang, Kai-Xun (China)
Zhang, Kewei (United Kingdom)
Zhang, Lei (China)
Zhang, Li Hua (China)
Zhang, Liangcai (China)
Zhang, Linbo (China)
Zhang, Linghai (USA)
Zhang, Liqun (China)
Zhang, Lixin (China)
Zhang, Lunchuan (China)
Zhang, Mei (China)
Zhang, Mengping (China)
Zhang, Ming Yi (China)
Zhang, Minghu (China)
Zhang, Na (China)
Zhang, Nansong (China)
Zhang, Peng (China)
Zhang, Ping (China)
Zhang, Ping (Turkey)

Zhang, Pingwen (China)
Zhang, Pu (China)
Zhang, Pu (Zhejiang) (China)
Zhang, Qi S. (USA)
Zhang, Qifan (China)
Zhang, Qingcai (China)
Zhang, Qingdong (China)
Zhang, Qingzheng (China)
Zhang, Qinhai (China)
Zhang, Runchu (China)
Zhang, Shaohua (China)
Zhang, Shaoyi (China)
Zhang, Shiqing (China)
Zhang, Shizhen (China)
Zhang, Shouchuan (China)
Zhang, Shugong (China)
Zhang, Shun (China)
Zhang, Shunyan (China)
Zhang, Sucheng (China)
Zhang, Tong (China)
Zhang, Weiping (China)
Zhang, Weiyi (China)
Zhang, Wenbo (China)
Zhang, Wenling (China)
Zhang, Wenqiong (China)
Zhang, Wensheng (China)
Zhang, Wentian (China)
Zhang, Xiang (China)
Zhang, Xiang Sun (China)
Zhang, Xianke (China)
Zhang, Xiao (China)
Zhang, Xiao Qiang (China)
Zhang, Xiao-Dong (China)
Zhang, Xiaolan (China)
Zhang, Xiaomin (China)
Zhang, Xiaowei (China)
Zhang, Xiaoxia (China)
Zhang, Xiaoyan (China)
Zhang, Xin (China)
Zhang, Xin Sheng (China)
Zhang, Xingyou (China)
Zhang, Xinli (China)
Zhang, Xinmiao (China)
Zhang, Xinting (China)
Zhang, Xi-Wen (China)
Zhang, Xuejuan (China)
Zhang, Xueyuan (China)
Zhang, Yan (China)
Zhang, Yanfen (China)
Zhang, Yanhong (China)
Zhang, Yanshuo (China)
Zhang, Yi (China)
Zhang, Yi (Zhejiang) (China)
Zhang, Yingbo (China)
Zhang, Yinglan (China)
Zhang, Yong (Canada)
Zhang, Yong (China)
Zhang, Yongbing (China)
Zhang, Yongli (Japan)
Zhang, Yonglin (China)
Zhang, Yongqian (China)
Zhang, Yuan (China)
Zhang, Yuehui (China)
Zhang, Yufeng (China)
Zhang, Yuhui (China)
Zhang, Yunfeng (Japan)
Zhang, Yuntao (China)
Zhang, Yuping (China)
Zhang, Yuqin (China)
Zhang, Zengxi (China)
Zhang, Zeyin (China)
Zhang, Zhanliang (China)
Zhang, Zhao (China)
Zhang, Zhenfeng (China)
Zhang, Zhenxiang (China)
Zhang, Zhifen (China)
Zhang, Zhikai (China)
Zhang, Zhiqiang (China)
Zhang, Zhitao (China)
Zhang, Zhixue (China)
Zhang, Zhiyue (China)
Zhang, Zhongfu (China)
Zhang, Zhong-Zhan (China)
Zhang, Zifang (China)
Zhanlav, Tugal (Mozambique)
Zhao, Bin (China)
Zhao, Chang-Jian (China)
Zhao, Chunlai (China)
Zhao, Deke (China)
Zhao, Di (China)

Zhao, Dong (China)
Zhao, Dongning (USA)
Zhao, Feng (China)
Zhao, Gaihuan (China)
Zhao, Huaizhong (United Kingdom)
Zhao, Huan-Xi (China)
Zhao, Huijiang (China)
Zhao, Jianli (China)
Zhao, Jianqiang (USA)
Zhao, Jingyu (China)
Zhao, Jinling (China)
Zhao, Juan (China)
Zhao, Kaiming (China)
Zhao, Liang (China)
Zhao, Lifeng (China)
Zhao, Linde (China)
Zhao, Ling (China)
Zhao, Linlong (China)
Zhao, Longhua (China)
Zhao, Meihua (China)
Zhao, Ou (China)
Zhao, Shufeng (China)
Zhao, Sufeng (China)
Zhao, Wenqiang (China)
Zhao, Wenzheng (China)
Zhao, Xiaohua (China)
Zhao, Xu An (China)
Zhao, Xuelei (China)
Zhao, Xuezhi (China)
Zhao, Yang (China)
Zhao, Yaoqing (China)
Zhao, Ying (China)
Zhao, Yongqiang (China)
Zhao, Yuanmin (China)
Zhao, Yuanyuan (China)
Zhao, Yude (Canada)
Zhao, Yufeng (China)
Zhao, Zhengang (China)
Zhao, Zhiyong (China)
Zhdanov, Alexander I. (Russia)
Zhen, Qiang (China)
Zheng, Chaozhou (China)
Zheng, Chongyou (China)
Zheng, Dechao (USA)
Zheng, Hao (China)
Zheng, Jianhua (China)
Zheng, Qibing (China)
Zheng, Quan (China)
Zheng, Quan (Wuhan)(China)
Zheng, Quan (Beijing) (China)
Zheng, Sining (China)
Zheng, Suwen (China)
Zheng, Weiying (China)
Zheng, Xinghua (China)
Zheng, Xiyin (China)
Zheng, Xuanyuan (China)
Zheng, Xue'an (China)
Zheng, Ye-Long (China)
Zheng, Yongai (China)
Zheng, Yongfan (China)
Zheng, Yu (China)
Zheng, Yumei (China)
Zheng, Yuxi (USA)
Zheng, Zhiming (China)
Zheng, Zhiyong (China)
Zheng, Zitu (China)
Zheng, Zuo-Huan (China)
Zhi, Lihong (China)
Zhislin, Grigorii M. (Russia)
Zhizhchenko, Alexey (Russia)
Zhong, Chengkui (China)
Zhong, Deshou (China)
Zhong, Huaijie (China)
Zhong, Shounan (China)
Zhong, Xiao (China)
Zhong, Yuquan (China)
Zhou, Aihui (China)
Zhou, Chunqin (China)
Zhou, Congyi (China)
Zhou, Denglin (China)
Zhou, Ding-Xuan (China)
Zhou, Fang (China)
Zhou, Fangmin (China)
Zhou, Feng (China)
Zhou, Guobiao (China)
Zhou, Haigang (China)
Zhou, Haiyun (China)
Zhou, Haizhong (China)
Zhou, Hong (China)
Zhou, Huan-Song (China)

Zhou, Huibin (USA)
Zhou, Jian (China)
Zhou, Jin (China)
Zhou, Jun (China)
Zhou, Junhua (China)
Zhou, Kouhua (Hefei)(China)
Zhou, Kouhua (China)
Zhou, Lingjun (China)
Zhou, Liren (China)
Zhou, Meng (China)
Zhou, Qing (China)
Zhou, Qinghua (China)
Zhou, Shengfan (China)
Zhou, Shulin (China)
Zhou, Taoguo (China)
Zhou, Wei (China)
Zhou, Xiang (China)
Zhou, Xiangyu (China)
Zhou, Xiaowen (Canada)
Zhou, Xinghe (China)
Zhou, Xunwei (China)
Zhou, Yifan (China)
Zhou, Yong (China)
Zhou, Yulin (China)
Zhou, Zehua (China)
Zhou, Zhan (China)
Zhou, Zhengxin (China)
Zhou, Zhenrong (China)
Zhou, Zixiang (China)
Zhu, Bin (China)
Zhu, Changrong (China)
Zhu, Chaofeng (China)
Zhu, Chenchang (USA)
Zhu, Chuanxi (China)
Zhu, Chunhao (China)
Zhu, Daoyuan (China)
Zhu, Deming (China)
Zhu, Dongjin (China)
Zhu, Fuhai (China)
Zhu, Fuliu (China)
Zhu, Guangtian (China)
Zhu, Huiyan (China)
Zhu, Jialin (China)
Zhu, Jiang (China)
Zhu, Jianping (China)
Zhu, Lin (China)
Zhu, Linsheng (China)
Zhu, Lixing (China)
Zhu, Meijun (USA)
Zhu, Ping (China)
Zhu, Qiding (China)
Zhu, Qingfeng (China)
Zhu, Shao Hong (China)
Zhu, Shenglin (China)
Zhu, Shixin (China)
Zhu, Tonglin (China)
Zhu, Wei (USA)
Zhu, Xiaohua (China)
Zhu, Xiaosheng (China)
Zhu, Xiping (China)
Zhu, Xuehong (China)
Zhu, Yanan (China)
Zhu, Yonghua (China)
Zhu, Yunmin (China)
Zhu, Zuo Nong (China)
Zhuang, Dongping (China)
Zhuang, Wei (China)
Zhuo, Jiliang (China)
Zhuravlev, S. G. (Russia)
Zhuravlev, Sergei G. (Russia)
Ziegenbalg, Jochen (Germany)
Ziegler, Guenter M. (Germany)
Zieschang, Paul-Hermann (USA)
Zimmermann, Bernd (Germany)
Zintl, Joerg (Germany)
Zoch, Avery S. (USA)
Zokayi, Ali Reza (Iran)
Zong, Chuanming (China)
Zorboska, Nina (Canada)
Zou, Jiancheng (China)
Zou, Jiezhong (China)
Zou, Xiangqun (China)
Zou, Xiong (China)
Zsidó, László (Italy)
Zuk, Andrzei (USA)
Zuo, Ling (China)
Zvyagin, Viktor G. (Russia)
Zworski, Maciej (USA)

ICM 2002 · Vol. I · 87–??

Participants by Country and Area

(according to their mailing address)

Algeria	3
Argentina	6
Armenia	3
Australia	28
Austria	11
Azerbaijan	1
Bahrain	1
Bangladesh	1
Barbados	1
Belarus	9
Belgium	10
Benin	1
Bosnia and Herzegovina	3
Botswana	3
Brazil	38
Brunei	1
Bulgaria	2
Cameroon	2
Canada	90
Chile	7
China	1973
Colombia	6
Croatia	5
Cuba	4
Czech	8
Denmark	9
DPR of Korea	7
Ecuador	1
Egypt	12
Estonia	5
Finland	22
France	94
Georgia	12
Germany	107
Ghana	1
Greece	5
Hungary	7
India	65
Indonesia	3
Iran	67
Ireland	3
Israel	31
Italy	58
Ivory Coast	2
Japan	191
Jordan	1
Kazakhstan	8
Kenya	1
Korea	95
Kyrgyzstan	4
Latvia	1
Lebanon	1
Lithuania	2
Luxembourg	1
Macedonia	1
Malawi	2
Malaysia	7
Mexico	31
Moldova	4
Mongolia	5
Morocco	8
Mozambique	5
Nepal	2
Netherlands	8
New Zealand	5
Nigeria	5
Norway	8
Oman	3
Pakistan	2
Peru	1
Philippines	12
Poland	26

Portugal 3
Qatar 1
Romania 14
Russia 167
Saudi Arabia 5
Senegal 1
Singapore 7
Slovakia 2
Slovenia 1
South Africa 17
Spain 26
Sri Lanka 1
Sudan 1
Sweden 19
Switzerland 17
Taiwan 13
Tajikistan 5
Tanzania 1
Thailand 40
Tunesia 8
Turkey 17
Ukraine 25
United Arab Emirates 1
United Kingdom 75
Uruguay 3
USA 459
Uzbekistan 17
Venezuela 6
Vietnam 23
Yugoslavia 11
Zambia 1
Zimbabwe 1

The Work of the Fields Medalists and of the Rolf Nevanlinna Prize Winner

ICM 2002 · Vol. I · 91–97

The Work of Laurent Lafforgue

Gérard Laumon*

Laurent Lafforgue has been awarded the Fields Medal for his proof of the Langlands correspondence for the full linear groups GL_r ($r \geq 1$) over function fields.

What follows is a brief introduction to the Langlands correspondence and to Lafforgue's theorem.

1. The Langlands correspondence

A global field is either a number field, i.e. a finite extension of $\mathbb{Q}$, or a function field of characteristic $p > 0$ for some prime number p, i.e. a finite extension of $\mathbb{F}_p(t)$ where $\mathbb{F}_p$ is the finite field with p elements. The global fields constitute a primary object of study in number theory and arithmetic algebraic geometry.

The conjectural Langlands correspondence, which was first formulated by Robert Langlands in 1967 in a letter to André Weil, relates two fundamental objects which are naturally attached to a global field F:

- its Galois group $\mathrm{Gal}(\overline{F}/F)$, where $\overline{F}$ is an algebraic closure of F, or more accurately its motivic Galois group of F which is by definition the tannakian group of the tensor category of Grothendieck motives over F,
- the ring $\mathbb{A}$ of adèles of F, or more precisely the collection of Hilbert spaces $L^2(G(F)\backslash G(\mathbb{A}))$ for all reductive groups G over F.

Roughly speaking, for any (connected) reductive group G over F, Langlands introduced a dual group ${}^{\mathrm{L}}G = \widehat{G} \rtimes \mathrm{Gal}(\overline{F}/F)$, the connected component $\widehat{G}$ of which is the complex reductive group whose roots are the co-roots of G and vice versa. And he predicted that a large part of the spectral decomposition of the Hilbert space $L^2(G(F)\backslash G(\mathbb{A}))$, equipped with the action by right translations of $G(\mathbb{A})$, is governed by representations of the motivic Galois group of F with values in ${}^{\mathrm{L}}G$.

Of special importance is the group $G = \mathrm{GL}_r$, the Langlands dual of which is simply the direct product ${}^{\mathrm{L}}\mathrm{GL}_r = \mathrm{GL}_r(\mathbb{C}) \times \mathrm{Gal}(\overline{F}/F)$. Indeed, any complex reductive group $\widehat{G}$ may be embedded into $\mathrm{GL}_r(\mathbb{C})$ for some r.

*CNRS and Université Paris-Sud, UMR 8628, Mathématique, F-91405 Orsay Cedex, France, gerard.laumon@math.u-psud.fr

The particular case $G = \mathrm{GL}_1$ of the Langlands correspondence is the abelian class field theory of Teiji Takagi and Emil Artin which was developed in the 1920s as a wide extension of the quadratic reciprocity law.

The Langlands correspondence embodies a large part of number theory, arithmetic algebraic geometry and representation theory of Lie groups. Small progress made towards this conjectural correspondence had already amazing consequences, the most striking of them being the proof of Fermat's last theorem by Andrew Wiles. Famous conjectures, such as the Artin conjecture on L-functions and the Ramanujan-Petersson conjecture, would follow from the Langlands correspondence.

2. Lafforgue's main theorem

Over number fields, the Langlands correspondence in its full generality seems still to be out of reach. Even its precise formulation is very involved. In the function field case the situation is much better. Thanks to Lafforgue, the Langlands correspondence for $G = \mathrm{GL}_r$ is now completely understood.

From now on, F is a function field of characteristic $p > 0$. We also fix some auxiliary prime number $\ell \neq p$.

As Alexandre Grothendieck showed, any algebraic variety over F gives rise to ℓ-adic representations of $\mathrm{Gal}(\overline{F}/F)$ on its étale cohomology groups and the irreducible ℓ-adic representations of $\mathrm{Gal}(\overline{F}/F)$ are good substitutes for irreducible motives over F. Therefore, the Langlands correspondence may be nicely formulated using ℓ-adic representations.

Let r be a positive integer. On the one hand, we have the set $\mathcal{G}_r$ of isomorphism classes of rank r irreducible ℓ-adic representations of $\mathrm{Gal}(\overline{F}/F)$ the determinant of which is of finite order. To each $\sigma \in \mathcal{G}_r$, Grothendieck attached an Eulerian product $L(\sigma, s) = \prod_x L_x(\sigma, s)$ over all the places x of F, which is in fact a rational function of p^{-s} and which satisfies a functional equation of the form $L(\sigma, s) = \varepsilon(\sigma, s)L(\sigma^\vee, 1-s)$ where $\sigma^\vee$ is the contragredient representation of σ and $\varepsilon(\sigma, s)$ is some monomial in p^{-s}. If σ is unramified at a place x, we have

$$L_x(\sigma, s) = \prod_{i=1}^{r} \frac{1}{1 - z_i p^{-s \deg(x)}}$$

where $z_1, \ldots, z_r$ are the *Frobenius eigenvalues* of σ at x and $\deg(x)$ is the degree of the place x.

On the other hand, we have the set $\mathcal{A}_r$ of isomorphism classes of cuspidal automorphic representations of $\mathrm{GL}_r(\mathbb{A})$ the central character of which is of finite order. Thanks to Langlands' theory of Eisenstein series, they are the building blocks of the spectral decomposition of $L^2(\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A}))$. To each $\pi \in \mathcal{A}_r$, Roger Godement and Hervé Jacquet attached an Eulerian product $L(\pi, s) = \prod_x L_x(\pi, s)$ over all the places x of F, which is again a rational function of p^{-s}, satisfying a functional equation $L(\pi, s) = \varepsilon(\pi, s)L(\pi^\vee, 1 - s)$. If π is unramified at a place x, we have

$$L_x(\pi, s) = \prod_{i=1}^{r} \frac{1}{1 - z_i p^{-s \deg(x)}}$$

where $z_1, \ldots, z_r$ are called the *Hecke eigenvalues* of π at x.

Theorem (i) (The Langlands Conjecture) *There is a unique bijective correspondence $\pi \to \sigma(\pi)$, preserving L-functions in the sense that $L_x(\sigma(\pi), s) = L_x(\pi, s)$ for every place x, between $\mathcal{A}_r$ and $\mathcal{G}_r$.*

(ii) (The Ramanujan-Petersson Conjecture) *For any $\pi \in \mathcal{A}_r$ and for any place x of F where π is unramified, the Hecke eigenvalues $z_1, \ldots, z_r \in \mathbb{C}^\times$ of π at x are all of absolute value* 1.

(iii) (The Deligne Conjecture) *Any $\sigma \in \mathcal{G}_r$ is pure of weight zero, i.e. for any place x of F where σ is unramified, and for any field embedding $\iota : \overline{\mathbb{Q}}_\ell \hookrightarrow \mathbb{C}$, the images $\iota(z_1), \ldots, \iota(z_r)$ of the Frobenius eigenvalues of σ at x are all of absolute value* 1.

As I said earlier, in rank $r = 1$, the theorem is a reformulation of the abelian class field theory in the function field case. Indeed, the reciprocity law may be viewed as an injective homomorphism with dense image

$$F^\times \backslash \mathbb{A}^\times \to \mathrm{Gal}(\overline{F}/F)^{\mathrm{ab}}$$

from the idèle class group to the maximal abelian quotient of the Galois group.

In higher ranks r, the first breakthrough was made by Vladimir Drinfeld in the 1970s. Introducing the fundamental concept of shtuka, he proved the rank $r = 2$ case. It is a masterpiece for which, among others works, he was awarded the Fields Medal in 1990.

3. The strategy

The strategy that Lafforgue is following, and most of the geometric objets that he is using, are due to Drinfeld. However, the gap between the rank two case and the general case was so big that it took more than twenty years to fill it.

Lafforgue considers the ℓ-adic cohomology of the moduli stack of rank r Drinfeld shtukas (see the next section) as a representation of $\mathrm{GL}_r(\mathbb{A}) \times \mathrm{Gal}(\overline{F}/F) \times \mathrm{Gal}(\overline{F}/F)$. By comparing the Grothendieck-Lefschetz trace formula (for Hecke operators twisted by powers of Frobenius endomorphisms) with the Arthur-Selberg trace formula, he tries to isolate inside this representation a subquotient which decomposes as

$$\bigoplus_{\pi \in \mathcal{A}_r} \pi \otimes \sigma(\pi)^\vee \otimes \sigma(\pi).$$

Such a comparison of trace formulas was first made by Yasutaka Ihara in 1967 for modular curves over $\mathbb{Q}$. Since, it has been extensively used for Shimura varieties and Drinfeld modular varieties by Langlands, Robert Kottwitz and many others. There are two main difficulties to overcome to complete the comparison:

- to prove suitable cases of a combinatorial conjecture of Langlands and Diana Shelstad, which is known as the *Fundamental Lemma*,
- to compare the contribution of the "fixed points at infinity" in the Grothendieck-Lefschetz trace formula with the weighted orbital integrals of James Arthur which occur in the geometric side of the Arthur-Selberg trace formula.

For the moduli space of shtukas, the required cases of the Fundamental Lemma were proved by Drinfeld in the 1970s. So, only the second difficulty was remaining after Drinfeld had completed his proof of the rank 2 case. This is precisely the problem that Lafforgue has solved after seven years of very hard work. The proof has been published in three papers totalling about 600 pages.

4. Drinfeld shtukas

Let X be "the" smooth, projective and connected curve over $\mathbb{F}_p$ whose field of rational functions is F. It plays the role of the ring of integers of a number field. Its closed points are the places of F. For any such point x we have the completion F_x of F at x and its ring of integers $\mathcal{O}_x \subset F_x$.

Let $\mathcal{O} = \prod_x \mathcal{O}_x \subset \mathbb{A}$ be the maximal compact subring of the ring of adèles. Weil showed that the double coset space

$$\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A}) / \mathrm{GL}_r(\mathcal{O})$$

can be naturally identified with the set of isomorphism classes of rank r vector bundles on X.

Starting from this observation, with the goal of realizing a congruence relation between Hecke operators and Frobenius endomorphisms, Drinfeld defined a rank r *shtuka* over an arbitrary field k of characteristic p as a diagram

$$ {}^{\tau}\mathcal{E} \xrightarrow[\varphi]{\sim} \overbrace{\mathcal{E}'' \hookrightarrow \mathcal{E}' \hookleftarrow \mathcal{E}}^{\text{Hecke}} $$

where $\mathcal{E}$, $\mathcal{E}'$ and $\mathcal{E}'$ are rank r vector bundles on the curve X_k deduced from X by extending the scalars to k, where $\mathcal{E} \hookrightarrow \mathcal{E}'$ is an elementary upper modification of $\mathcal{E}$ at some k-rational point of X which is called the *pole* of the shtuka, where $\mathcal{E}'' \hookrightarrow \mathcal{E}'$ is an elementary lower modification of $\mathcal{E}'$ at some k-rational point of X which is called the *zero* of the shtuka, and where ${}^{\tau}\mathcal{E}$ is the pull-back of $\mathcal{E}$ by the endomorphism of X_k which is the identity on X and the Frobenius endomorphism on k.

Drinfeld proved that the above shtukas are the k-rational points of an algebraic stack over $\mathbb{F}_p$ which is equipped with a projection onto $X \times X$ given by the pole and the zero. More generally, he introduced level structures on rank r shtukas and he constructed an algebraic stack Sht_r parametrizing rank r shtukas equipped with a compatible system of level structures. This last algebraic stack is endowed with an algebraic action of $\mathrm{GL}_r(\mathbb{A})$ through the Hecke operators.

5. Iterated shtukas

The geometry at infinity of the moduli stack Sht_r is amazingly complicated. The algebraic stack Sht_r is not of finite type and one needs to *truncate* it to obtain manageable geometric objects. Bounding the Harder-Narasimhan polygon of a shtuka, Lafforgue defines a family of open substacks $(\mathrm{Sht}_r^{\leq P})_P$ which are all of finite type and whose union is the whole moduli stack. But in doing so, he loses the action of the Hecke operators which do not stabilize those open substacks.

In order to recover the action of the Hecke operators, Lafforgue enlarges Sht_r by allowing specific degenerations of shtukas that he has called *iterated shtukas.*

More precisely, Lafforgue lets the isomorphism $\varphi : {}^{\tau}\mathcal{E} \xrightarrow{\sim} \mathcal{E}''$ appearing in the definition of a shtuka, degenerate to a *complete homomorphism* ${}^{\tau}\mathcal{E} \Rightarrow \mathcal{E}''$, i.e. a continuous family of complete homomorphisms between the stalks of the vector bundles ${}^{\tau}\mathcal{E}$ and $\mathcal{E}''$.

Let me recall that a complete homomorphism $V \Rightarrow W$ between two vector spaces of the same dimension r is a point of the partial compactification $\widetilde{\mathrm{Hom}}(V,W)$ of $\mathrm{Isom}(V,W)$ which is obtained by successively blowing up the quasi-affine variety $\mathrm{Hom}(V,W) - \{0\}$ along its closed subsets

$$\{f \in \mathrm{Hom}(V,W) - \{0\} \mid \mathrm{rank}(f) \leq i\}$$

for $i = 1, \ldots, r-1$. If $V = W$ is the standard vector space of dimension r, the quotient of $\widetilde{\mathrm{Hom}}(V,W)$ by the action of the homotheties is the Procesi-De Concini compactification of PGL_r.

In particular, Lafforgue obtains a smooth compactification, with a normal crossing divisor at infinity, of any truncated moduli stack of shtukas without level structure.

6. One key of the proof

Lafforgue proves his main theorem by an elaborate induction on r. Compared to Drinfeld's proof of the rank 2 case, a very simple but crucial novelty in Lafforgue's proof is the distinction in the ℓ-adic cohomology of Sht_r between the r-negligible part (the part where all the irreducible constituents as Galois modules are of dimension $< r$) and the r-essential part (the rest). Lafforgue shows that the difference between the cohomology of Sht_r and the cohomology of any truncated stack $\mathrm{Sht}_r^{\leq P}$ is r-negligible. He also shows that the cohomology of the boundary of $\mathrm{Sht}_r^{\leq P}$ is r-negligible. Therefore, the r-essential part, which is defined purely by considering the Galois action and which is naturally endowed with an action of the Hecke operators, occurs in the ℓ-adic cohomology of any truncated moduli stack $\mathrm{Sht}_r^{\leq P}$ and also in their compactifications.

At this point, Lafforgue makes an extensive use of the proofs by Richard Pink and Kazuhiro Fujiwara of a conjecture of Deligne on the Grothendieck-Lefschetz trace formula.

7. Compactification of thin Schubert cells

In proving the Langlands conjecture for functions fields, Lafforgue tried to construct nice compactifications of the truncated moduli stacks of shtukas with arbitrary level structures. A natural way to do that is to start with some nice compactifications of the quotients of $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$ for all integers $n \geq 1$, and to apply a procedure similar to the one which leads to iterated shtukas.

Lafforgue constructed natural compactifications of $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$. In fact, he remarked that $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$ is the quotient of $\mathrm{GL}_r^{n+1} / \mathrm{GL}_r$ by the obvious free action of the torus $\mathbb{G}_{\mathrm{m}}^{n+1} / \mathbb{G}_{\mathrm{m}}$ and that $\mathrm{GL}_r^{n+1} / \mathrm{GL}_r$ may be viewed as a thin Schubert cell in the Grassmannian variety of r-planes in a $r(n+1)$-dimensional vector space. And, more generally, he constructed natural compactifications of all similar quotients of thin Schubert cells in the Grassmannian variety of r-planes in a finite-dimensional vector space.

Let me recall that thin Schubert cells are by definition intersections of Schubert varieties and that Israel Gelfand, Mark Goresky, Robert MacPherson and Vera Serganova constructed natural bijections between thin Schubert cells, matroids and certain convex polyhedra which are called polytope matroids.

For $n = 1$ and arbitrary r, Lafforgue's compactification of $\mathrm{PGL}_r^2 / \mathrm{PGL}_r$ coincides with the Procesi-De Concini compactification of PGL_r. It is smooth with a normal crossing divisor at infinity.

For $n = 2$ and arbitrary r, Lafforgue proves that his compactification of $\mathrm{PGL}_r^3 / \mathrm{PGL}_r$ is smooth over a toric stack, and thus can be desingularized.

For $n \geq 3$ and $r \geq 3$, the geometry of Lafforgue's compactifications is rather mysterious and not completely understood.

Gerd Faltings linked the search of good local models for Shimura varieties in bad characteristics to the search of smooth compactifications of G^{n+1}/G for a reductive group G. He gave another construction of Lafforgue's compactifications of $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$ and he succeeded in proving that Lafforgue's compactification of $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$ is smooth for $r = 2$ and arbitrary n.

8. Conclusion

I hope that I gave you some idea of the depth and the technical strength of Lafforgue's work on the Langlands correspondence for which we are now honoring him with the Fields Medal.

Laurent Lafforgue

I.H.É.S., Bures-sur-Yvette, France
Né le 6 novembre 1966 à Antony (Hauts-de-Seine), France
Nationalité française

1986–1990	Élève à l'École Normale Supérieure de Paris
1988–1991	Étudiant en géométrie algébrique (et en théorie d'Arakelov avec Christophe Soulé)
1990–1991 puis 1992–2000	Chargé de recherche au C.N.R.S. dans l'équipe "Arithmétique et Géométrie Algébrique" de l'Université Paris-Sud
1991–1992	Service militaire à l'École Spéciale Militaire de Saint-Cyr-Coëtquidan
1993–1994	Thèse sur les D-chtoucas de Drinfeld sous la direction de Gérard Laumon
1994–2000	Suite de l'étude des chtoucas
2000	Professeur à l'Institut des Hautes Études Scientifiques

L. Lafforgue (left) and G. Laumon

ICM 2002 · Vol. I · 99–103

The Work of Vladimir Voevodsky

Christophe Soulé*

Vladimir Voevodsky was born in 1966. He studied at Moscow State University and Harvard university. He is now Professor at the Institute for Advanced Study in Princeton.

Among his main achievements are the following: he defined and developed motivic cohomology and the $\mathbf{A}^1$-homotopy theory of algebraic varieties; he proved the Milnor conjectures on the K-theory of fields.

Let us state the first Milnor conjecture. Let F be a field and n a positive integer. The *Milnor K-group* of F is the abelian group $K_n^M(F)$ defined by the following generators and relations. The generators are sequences $\{a_1, \dots, a_n\}$ of n units $a_i \in F^*$. The relations are

$$\begin{aligned} & \{a_1, \dots, a_{k-1}, xy, a_{k+1}, \dots, a_n\} \\ = \; & \{a_1, \dots, a_{k-1}, x, a_{k+1}, \dots, a_n\} + \{a_1, \dots, a_{k-1}, y, a_{k+1}, \dots, a_n\} \end{aligned}$$

for all $a_i, x, y \in F^*$, $1 \leq k \leq n$, and the *Steinberg relation*

$$\{a_1, \dots, x, \dots, 1-x, \dots, a_n\} = 0$$

for all $a_i \in F^*$ and $x \in F - \{0, 1\}$.

On the other hand, let $\overline{F}$ be an algebraic closure of F and $G = \mathrm{Gal}(\overline{F}/F)$ the absolute Galois group of F, with its profinite topology. The *Galois cohomology* of F with $\mathbf{Z}/2$ coefficients is, by definition,

$$H^n(F, \mathbf{Z}/2) = H^n_{\text{continuous}}(G, \mathbf{Z}/2)\,.$$

Theorem 1. (Voevodsky 1996 [5]) *Assume $1/2 \in F$ and $n \geq 1$. The Galois symbol*

$$h_n : K_n^M(F)/2K_n^M(F) \to H^n(F, \mathbf{Z}/2)$$

is an isomorphism.

*CNRS and Institut des Hautes Etudes Scientifiques, 35, route de Chartres, 91440 Bures-sur-Yvette, France.

This was conjectured by Milnor in 1970 [1]. When $n = 2$, Theorem 1 was proved by Merkurjev in 1983. The case $n = 3$ was then solved independently by Merkurjev-Suslin and Rost.

There exists also a Galois symbol on $K_n^M(F)/p\,K_n^M(F)$ for any prime p invertible in F. When $n = 2$ and F is a number field, Tate proved that it is an isomorphism. In 1983 Merkurjev and Suslin proved that it is an isomorphism when $n = 2$ and F is any field. Both Voevodsky and Rost have made a lot of progress towards proving that, for any F, any $n > 0$ and any p invertible in F, the Galois symbol is an isomorphism.

The map h_n in Theorem 1 is defined as follows. When $n = 1$, we have $K_1^M(F) = F^*$ and $H^1(F, \mathbf{Z}/2) = \mathrm{Hom}(G, \mathbf{Z}/2)$. The map

$$h_1 : F^*/(F^*)^2 \to \mathrm{Hom}(G, \mathbf{Z}/2)$$

maps $a \in F^*$ to the quadratic character χ_a defined by

$$\chi_a(g) = g(\sqrt{a})/\sqrt{a} = \pm 1$$

for any $g \in G$ and any square root $\sqrt{a}$ of a in $\overline{F}$. That h_1 is bijective is a special case of Kummer theory. When $n \geq 2$, we just need to define h_n on the generators $\{a_1, \ldots, a_n\}$ of $K_n^M(F)$. It is given by a cup-product:

$$h_n(\{a_1, \ldots, a_n\}) = \chi_{a_1} \cup \cdots \cup \chi_{a_n}\,.$$

The fact that h_n is compatible with the Steinberg relation was first noticed by Bass and Tate.

Theorem 1 says that $H^n(F, \mathbf{Z}/2)$ has a very explicit description. In particular, an immediate consequence of Theorem 1 and the definition of h_n is the following

Corollary 1. *The graded* $\mathbf{Z}/2$*-algebra* $\bigoplus_{n \geq 0} H^n(F, \mathbf{Z}/2)$ *is spanned by elements of degree one.*

This means that absolute Galois groups are very special groups. Indeed, it is seldom seen that the cohomology of a group or a topological space is spanned in degree one.

Corollary 2. (Bloch) *Let X be a complex algebraic variety and $\alpha \in H^n(X(\mathbf{C}), \mathbf{Z})$ a class in its singular cohomology. Assume that $2\alpha = 0$. Then, there exists a nonempty Zariski open subset $U \subset X$ such that the restriction of α to U vanishes.*

If Theorem 1 was extended to $K_n^M(F)/p\,K_n^M(F)$ for all n and p, Corollary 2 would say that any torsion class in the integral singular cohomology of X is supported on some hypersurface. (Hodge seems to have believed that such a torsion class should be Poincaré dual to an analytic cycle, but this is not always true.)

With Orlov and Vishik, Voevodsky proved a second conjecture of Milnor relating the Witt group of quadratic forms over F to its Milnor K-theory [3].

A very serious difficulty that Voevodsky had to overcome to prove Theorem 1 was that, when $n = 2$, Merkurjev made use of the algebraic K-theory of conics over F, but, when $n \geq 2$, one needed to study special quadric hypersurfaces of dimension $2^{n-1} - 1$. And it is quite hard to compute the algebraic K-theory of varieties of such a high dimension. Although Rost had obtained crucial information about the K-theory of these quadrics, this was not enough to conclude the proof when $n > 3$. Instead of algebraic K-theory, Voevodsky used *motivic cohomology*, which turned out to be more computable.

Given an algebraic variety X over F and two integers $p, q \in \mathbf{Z}$, Voevodsky defined an abelian group $H^{p,q}(X, \mathbf{Z})$, called motivic cohomology. These groups are analogs of the singular cohomology of CW-complexes. They satisfy a long list of properties, which had been anticipated by Beilinson and Lichtenbaum. For example, when n is a positive integer and X is smooth, the group

$$H^{2n,n}(X, \mathbf{Z}) = \mathrm{CH}^n(X)$$

is the Chow group of codimension n algebraic cycles on X modulo linear equivalence. And when X is a point we have

$$H^{n,n}(\text{point}) = K_n^M(F)\,.$$

It is also possible to compute Quillen's algebraic K-theory from motivic cohomology. Earlier constructions of motivic cohomology are due to Bloch (at the end of the seventies) and, later, to Suslin. The way Suslin modified Bloch's definition was crucial to Voevodsky's approach and, as a matter of fact, several important papers on this topic were written jointly by Suslin and Voevodsky [4, 7]. There exist also two very different definitions of $H^{p,q}(X, \mathbf{Z})$, due to Levine and Hanamura; according to the experts they lead to the same groups. But it seems fair to say that Voevodsky's approach to motivic cohomology is the most complete and satisfactory one.

A larger context in which Voevodsky developed motivic cohomology is the $\mathbf{A}^1$-homotopy of algebraic manifolds [6], which is a theory of "algebraic varieties up to deformations", developed jointly with Morel [2]. Starting with the category of smooth manifolds (over a fixed field F), they first embed this category into the category of *Nisnevich sheaves*, by sending a given manifold to the sheaf it represents. A Nisnevich sheaf is a sheaf of sets on the category of smooth manifolds for the Nisnevich topology, a topology which is finer (resp. coarser) than the Zariski (resp. étale) topology. Then Morel and Voevodsky define a homotopy theory of Nisnevich sheaves in much the same way the homotopy theory of CW-complexes is defined. The parameter space of deformations is the affine line $\mathbf{A}^1$ instead of the real unit interval $[0, 1]$. Note that, in this theory there are *two circles* (corresponding to the two degrees p and q for motivic cohomology)! The first circle is the sheaf represented by the smooth manifold $\mathbf{A}^1 - \{0\}$ (indeed, $\mathbf{C} - \{0\}$ has the homotopy type of a circle). The second circle is $\mathbf{A}^1/\{0, 1\}$ (note that $\mathbf{R}/\{0, 1\}$ is a loop). The latter is not represented by a smooth manifold. But, if we identify 0 and 1 in the sheaf

of sets represented by $\mathbf{A}^1$ we get a presheaf of sets, and $\mathbf{A}^1/\{0,1\}$ can be defined as the sheaf attached to this presheaf. This example shows why it was useful to embed the category of algebraic manifolds into a category of sheaves.

It is quite extraordinary that such a homotopy theory of algebraic manifolds exists at all. In the fifties and sixties, interesting invariants of differentiable manifolds were introduced using algebraic topology. But very few mathematicians anticipated that these "soft" methods would ever be successful for algebraic manifolds. It seems now that any notion in algebraic topology will find a partner in algebraic geometry. This has long been the case with Quillen's algebraic K-theory, which is precisely analogous to topological K-theory. We mentioned that motivic cohomology is an algebraic analog of singular cohomology. Voevodsky also computed the algebraic analog of the Steenrod algebra, i.e. cohomological operations on motivic cohomology (this played a decisive role in the proof of Theorem 1). Morel and Voevodsky developed the (stable) $\mathbf{A}^1$-homotopy theory of algebraic manifolds. Voevodsky defined *algebraic cobordism* as homotopy classes of maps from the suspension of an algebraic manifold to the classifying space MGL. There is also a direct geometric definition of algebraic cobordism, due to Levine and Morel (see Levine's talk in these proceedings), which should compare well with Voevodsky's definition. And the list is growing: Morava K-theories, stable homotopy groups of spheres, etc...

Vladimir Voevodsky is an amazing mathematician. He has demonstrated an exceptional talent for creating new abstract theories, about which he proved highly nontrivial theorems. He was able to use these theories to solve several of the main long standing problems in algebraic K-theory. The field is completely different after his work. He opened large new avenues and, to use the same word as Laumon, he is leading us closer to the world of *motives* that Grothendieck was dreaming about in the sixties.

References

[1] John Milnor, Algebraic K-theory and quadratic forms, *Inv. Math.*, 9 (1970), 318–344.

[2] Fabien Morel & Vladimir Voevodsky, $\mathbf{A}^1$-homotopy theory of schemes, *Publ. Math. IHES*, 90 (1999), 45–143.

[3] D. Orlov, A. Vishik & Vladimir Voevodsky, An exact sequence for Milnor's K-theory with applications to quadratic forms (2000), to appear.

[4] Andrei Suslin & Vladimir Voevodsky, Singular homology of abstract algebraic varieties, *Inv. Math.*, 123 (1996), 61–94.

[5] Vladimir Voevodsky, On 2-torsion in motivic cohomology (2001), to appear.

[6] Vladimir Voevodsky, The $\mathbf{A}^1$-homotopy theory. In *Proceedings of the international congress of mathematians*, Volume 1, pp. 579–604, Berlin, 1998.

[7] Vladimir Voevodsky, Andrei Suslin & Eric Friedlander, Cycles, transfers and motivic homology theories, *Annals of Maths. Studies* 143, Princeton University Press (2000).

Vladimir Voevodsky

June 4, 1966

1989	B.S. in Mathematics, Moscow University
1992	Ph.D. in Mathematics, Harvard University
1992–1993	Institute for Advanced Study, Member
1993–1996	Harvard University, Junior Fellow of Harvard Society of Fellows
1996–1997	Harvard University, Visiting Scholar
1996–1997	Max-Planck Institute, Visiting Scholar
1996–1999	Northwestern University, Associate Professor
1998–2001	Institute for Advanced Study, Member
2002	Institute for Advanced Study, Professor

V. Voevodsky (right) and C. Soulé

ICM 2002 · Vol. I · 105–115

On the Work of Madhu Sudan: the 2002 Nevalinna Prize Winner

Shafi Goldwasser*

1. Introduction

Madhu Sudan's work spans many areas of computer science theory including computational complexity theory, the design of efficient algorithms, algorithmic coding theory, and the theory of program checking and correcting.

Two results of Sudan stand out in the impact they have had on the mathematics of computation. The first work shows a probabilistic characterization of the class NP – those sets for which short and easily checkable proofs of membership exist, and demonstrates consequences of this characterization to classifying the complexity of approximation problems. The second work shows a polynomial time algorithm for list decoding the Reed Solomon error correcting codes.

This short note will be devoted to describing Sudan's work on probabilistically checkable proofs – the so called *PCP theorem* and its implications. We refer the reader to [29, 30] for excellent expositions on Sudan's breakthrough work on list decoding, and its impact on the study of computational aspects of coding theory as well as the use of coding theory within complexity theory.

Complexity theory is concerned with how many resources such as time and space are required to perform various computational tasks. Computational tasks arise in classical mathematics as well as in the world of computer science and engineering. Examples of what we may call a computational task include finding a proof for a mathematical theorem, automatic verification of the correctness of a given mathematical proof, and designing algorithms for transmitting information reliably through a noisy channel of communication. Defining what is a 'success' when solving some of these computational tasks is still a lively and important part of research in this stage of development of complexity theory.

A large body of Sudan's work, started while he was working on his PhD thesis, addresses the automatic verification of the correctness of mathematical proofs. Many issues come up: how should we encode a mathematical proof so that a computer can verify it, which mathematical statements have proofs which can be quickly verified, and what is the relation between the size of the description of the theorem

*Weizmann Institute of Science, Israel and Massachusetts Institute of Technology, USA

and the size of its shortest proof which can be quickly verified. The work of Sudan sheds light on all of these questions.

2. Efficient proof checking

Let us start with the classic notion of efficiently checkable proofs, which goes back to the early days of computer science in the early seventies when the NP class was defined [8, 25].

Definition 1 *The class NP consists of those sets $L \subset \{0,1\}^*$ for which there exists polynomial time verification algorithm V_L and polynomial p such that $x \in L$ if and only if there exists a $y_x \in \{0,1\}^{p(|x|)}$ which makes $V_L(x, y_x) = TRUE$. We call V_L the NP-verifier for the language $L \in NP$, and y_x the NP-witness for x in L.*

One example of $L \in NP$ is the set of pairs (G, k) where $k \in \mathbf{Z}$ and G is a graph which contain a complete subgraph on k vertices – the so called CLIQUE problem. The NP-witness for pair $(G, k) \in CLIQUE$ is the complete subgraph in G of size k. Another example is the set of all logical formulas for which a truth assignment to its Boolean variables exists which makes it true – the SATISFIABILITY problem. The NP-witness for a logical formula ϕ is a particular setting of its variables which make the formula satisfiable. Graphs, logical formulas, and truth assignments can all be encoded as binary strings.

3. Probabilistic checking of proofs

In the eighties, extensions of the notion of an efficiently verifiable proof were proposed to address issues arising in disciplines involving interactive computation such as cryptography. The extensions incorporate the idea of using randomness in the verification process and allow a negligible probability of error to be present in the verification process. Variants of probabilistic proof systems include interactive proofs [16], public-coin interactive proofs [3], computational arguments[4], CS-proofs [26], Holographic proofs [6], multi-prover interactive proofs [7], memoryless oracles [14], and probabilistically checkable proofs [10, 2]. The latter three definitions are equivalent to each other although each was introduced under a different name.

By the early nineties probabilistically checkable proofs proofs were generally accepted as the right extension for complexity theoretic investigations. The class PCP of sets for which membership can be checked by "probabilistically checkable proofs" is defined as follows.

Definition 2 *Let $L \subset \{0,1\}^*$. For L in PCP, there exists a probabilistic polynomial time verification algorithm V_L*

- *if $x \in L$, then there exists a $O_x \in \{0,1\}^*$ such that $Prob[V_L^{O_x}(x) = TRUE] > 1$*
- *if $x \notin L$, then for all $O_x \in \{0,1\}^*$, $Prob[V_L^{O_x}(x) = TRUE] < \frac{1}{2}$.*

The probabilities above are taken over the probabilistic choices of the verification algorithm V_L. The notation $V_L^{O_x}$ means that V_L does not receive O_x as an input but rather can read individual bits in O_x by specifying their locations explicitly. We call V_L the PCP-verifier for $L \in PCP$, and O_x the PCP-witness for x in L.

A few comments are in order.

For each bit of O_x read, we charge V_L for the time it takes to write down the address of the bit to be read. The requirement that V_L runs in polynomial time implies then that the length of the PCP-witness for x is bounded by an exponential in $|x|$.

A verifier may make an error and accept incorrectly, but the probability of this event can be made exponentially (in $|x|$) small by running a polynomial number of independent executions of V_L and accepting only if all executions accept. In light of the above, we argue that probabilistically checkable proofs capture what we want from any efficiently checkable proof system: correct statements are always accepted, incorrect statement are (almost) never accepted, and the verification procedure terminates quickly.

Are probabilistically checkable proofs more powerful than the deterministic NP style proofs? Developments made in a sequence of beautiful papers [32, 24, 5], finally culminated in the result of Babai et. al. [5] showing that indeed $PCP = NEXPTIME$.[1] By the separation of the non-deterministic time hierarchy, it is known that NP is strictly contained in $NEXPTIME$. Thus indeed, the probabilistic checking of proofs is more powerful than the classical deterministic one (at least when the verifier is restricted to polynomial time).

Soon after the power of PCP verifiers was characterized, a finer look was taken at the resources PCP verifiers use. Two important resources in classifying the complexity of language L were singled out [10]: the amount of randomness used by the PCP verifier and the number of bits it reads from the PCP-witness (the latter number is referred to as the *query size* of V_L).

Definition 3 *Let $PCP(r(n), q(n))$ denote class of sets $L \in PCP$ for which there exists a PCP verifier for L which on input $x \in \{0,1\}^n$ uses at most $O(r(n))$ random bits and reads at most $O(q(n))$ bits of the witness oracle O_x.* [2]

Obviously, $NP \subset PCP(0, \cup_c n^c)$ as an NP verifier is simply a special case of the PCP verifier which does not use any randomness. Starting with scaling down the result of [5] it was shown (or at least implied) in a sequence of improvements [6, 10, 2] that $NP \subset PCP(\log n, poly(\log n))$. These results successively lowered the number of bits that the PCP- verifier needs to read from the PCP-witness, but it seemed essential for the correctness of the verification procedure that this number should be a function which grows with the size of the input.

In the eighties, extensions of the notion of an efficiently verifiable proof were proposed to address issues arising in disciplines involving interactive computation

[1] The class $NEXPTIME$ is defined exactly in the same manner as NP except that the verifier V_L has exponential time and the witness may be exponentially long.

[2] $O(g(n) = cf(n)$ s.t. there exists a constant c such that $g(n) \leq cf(n)$ for all n sufficiently large}

such as cryptography. The extensions incorporate the idea of using randomness in the verification process and allow a negligible probability of error to be present in the verification process. Variants of probabilistic proof systems include interactive proofs [16], public-coin interactive proofs [3], computational arguments[4], CS-proofs [26], Holographic proofs [6], multi-prover interactive proofs [7], memoryless oracles [14], and probabilistically checkable proofs [10, 2]. The latter three definitions are equivalent to each other although each was introduced under a different name.

By the early nineties probabilistically checkable proofs proofs) were generally accepted as the right extension for complexity theoretic investigations. The class PCP of sets for which membership can be checked by "probabilistically checkable proofs" is defined as follows.

Definition 4 *Let $L \subset \{0,1\}^*$. For L in PCP, there exists a probabilistic polynomial time verification algorithm V_L*

- *if $x \in L$, then there exists a $O_x \in \{0,1\}^*$ such that $Prob[V_L^{O_x}(x) = TRUE] > 1$*
- *if $x \notin L$, then for all $O_x \in \{0,1\}^*$, $Prob[V_L^{O_x}(x) = TRUE] < \frac{1}{2}$.*

The probabilities above are taken over the probabilistic choices of the verification algorithm V_L. The notation $V_L^{O_x}$ means that V_L does not receive O_x as an input but rather can read individual bits in O_x by specifying their locations explicitly. We call V_L the PCP-verifier for $L \in PCP$, and O_x the PCP-witness for x in L.

A few comments are in order.

For each bit of O_x read, we charge V_L for the time it takes to write down the address of the bit to be read. The requirement that V_L runs in polynomial time implies then that the length of the PCP-witness for x is bounded by an exponential in $|x|$.

A verifier may make an error and accept incorrectly, but the probability of this event can be made exponentially (in $|x|$) small by running a polynomial number of independent executions of V_L and accepting only if all executions accept. In light of the above, we argue that probabilistically checkable proofs capture what we want from any efficiently checkable proof system: correct statements are always accepted, incorrect statement are (almost) never accepted, and the verification procedure terminates quickly.

Are probabilistically checkable proofs more powerful than the deterministic NP style proofs? Developments made in a sequence of beautiful papers [32, 24, 5], finally culminated in the result of Babai et. al. [5] showing that indeed $PCP = NEXPTIME$.[3] By the separation of the non-deterministic time hierarchy, it is known that NP is strictly contained in $NEXPTIME$. Thus indeed, the probabilistic checking of proofs is more powerful than the classical deterministic one (at least when the verifier is restricted to polynomial time).

[3]The class $NEXPTIME$ is defined exactly in the same manner as NP except that the verifier V_L has exponential time and the witness may be exponentially long.

Soon after the power of PCP verifiers was characterized, a finer look was taken at the resources PCP verifiers use. Two important resources in classifying the complexity of language L were singled out [10]: the amount of randomness used by the PCP verifier and the number of bits it reads from the PCP-witness (the latter number is referred to as the *query size* of V_L).

Definition 5 *Let $PCP(r(n), q(n))$ denote class of sets $L \in PCP$ for which there exists a PCP verifier for L which on input $x \in \{0,1\}^n$ uses at most $O(r(n))$ random bits and reads at most $O(q(n))$ bits of the witness oracle O_x.* [4]

Obviously, $NP \subset PCP(0, \cup_c n^c)$ as an NP verifier is simply a special case of the PCP verifier which does not use any randomness. Starting with scaling down the result of [5] it was shown (or at least implied) in a sequence of improvements [6, 10, 2] that $NP \subset PCP(\log n, poly(\log n))$. These results successively lowered the number of bits that the PCP verifier needs to read from the PCP-witness, but it seemed essential for the correctness of the verification procedure that this number should be a function which grows with the size of the input.

4. The PCP theorem

In a breakthrough, which has since become known as *the PCP theorem*, Sudan and his co-authors characterized the class NP exactly in terms of PCP. They showed that NP contains exactly those languages in which a PCP-verifier can verify membership using *only* a constant query size and using logarithmic (in the instance size) number of coins. More over, there exists a polynomial time procedure to transform an NP-witness of x in L into a PCP-witness of x in L.

Theorem 6 *[1] $NP = PCP(\log n, 1)$*

On an intuitive level, the PCP theorem says that there exist a probabilistic verifier for proofs of mathematical assertions which can look only at a constant number of bit positions at the proof and yet with some positive probability catch any mistake made in a fallacious argument.

The proof of the PCP theorem is deep, beautiful, and quite complex. It brings together ideas from algebra, error correcting codes, probabilistic computation, and program testing.

Although the PCP theorem establishes a complexity result, its proof is algorithmic in nature, as it is a transformation of an NP-witness and a deterministic NP-verifier for $L \in NP$ into a PCP-witness and an PCP-verifier for L. As such it uses methods from the design of computer algorithms and the design of error correcting codes. Several excellent expositions of the proof appeared [28].

In a very strong sense, the act of transforming an NP witness into a PCP witness is similar to transforming a message into an error correcting code word. The analogy being that a code word is an encoding of a message which enables

[4] $O(g(n) = cf(n)$ s.t. there exists a constant c such that $g(n) \leq cf(n)$ for all n sufficiently large}

error detection in spite of noise injected by an adversary, and a PCP witness is an encoding of a proof which enables detection with high probability of an error in spite the best efforts to hide it made by a cheating pretend-to-be prover.

Yet, the act of classic *decoding* of a code word is very different than the act of checking the correctness of a PCP witness . Whereas in error correcting codes one attempts to recover the entire original message from the corrupted code word if too much noise has not occurred; here we only want to verify that the PCP-witness is a proper encoding of a valid NP-witness (of the same fact) which would have convinced an NP-verifier to accept. It suffices to read only a constant number of bit positions to achieve the latter task, whereas the decoding task depends on reading the entire code word.

One of the subsequent contributions of Sudan, involves constructing a new type of *locally testable codes* [11, 17]. Locally testable codes are error-correcting codes for which error detection can be made with probability proportional to the distance of the non-codeword from the code, based on reading only a constant number of (random) symbols from the received word. A related concept is that of *locally decodable codes* [23, 18] which are error correcting codes which may enable recovery of part of the message (rather than the entire message) by reading only part of the corrupted code word.

5. PCP and hardness of approximation

The intellectual appeal of the PCP theorem statement is obvious. What is much less obvious and what has been the main impact of the PCP theorem is its usefulness in proving NP hardness of many approximate versions of combinatorial optimization problems. A task which alluded the theoretical computer science community for over twenty years.

Shortly after the class NP and the companion notion of an NP-complete and NP-hard problems[5] were introduced, Karp illustrated its great relevance to combinatorial optimization problems in his 1974 paper [22], He showed that a wide collection of optimization problems (including the minimum travelling salesman problem in a graph, integer programming, minimum graph coloring and maximum graph clique suitably reformulated as language membership problems) are NP-complete. Proving that a problem is NP-complete is generally taken to mean that they are intrinsically intractable as otherwise $NP = P$.

In practice this means there is no point in wasting time trying to devise efficient algorithms for NP-complete problems, as none exists (again if $NP \neq P$). Still these problems do come up in applications all the time, and need to be addressed. The question is, how? Several methods for dealing with NP-completeness arose in the last 20 years.

One technique is to devise algorithms which provably work efficiently for particular input distribution on the instances ("average" instances) of the NP-complete

[5] A set is $\mathcal{NP}$-hard if *any* efficient algorithm for it, can be used to efficiently decide every other set in $\mathcal{NP}$. An NP set which is NP-hard is called NP-complete. By definition, every $\mathcal{NP}$-complete language is as hard to compute as any other.

problems. It is not clear however how to determine whether your application produces such input distribution.

Another direction has been to devise approximation algorithms. We say that an approximation algorithm α-approximates a maximization problem if, for every instance, it provably guarantees a solution of value which is at least $\frac{1}{\alpha}$ of the value of an optimal solution; an approximation algorithm is said to α-approximate a minimization problem if it guarantees a solution of value at most α of the value of an optimal solution.

Devising approximation algorithms has been an active research area for twenty years, still for many NP-hard problems success has been illusive whereas for others good approximation factors were achievable. There has been no theoretical explanation of this state of affairs. Attempting to prove that approximating the solution to NP-hard problems is in itself NP-hard were not successful.

The PCP theorems of Sudan and others, starting with the work of Feige et. al. [10], has completely revolutionized this state of affairs. It is now possible using the PCP characterization of NP to prove that approximating many optimization problems each for different approximation factors is in itself NP-hard. The mysteries of why it is not only hard to solve optimization problems exactly but also approximately, and why different NP-hard problems behave differently with respect to approximation have been resolved.

The connection between bounding the randomness and query complexity of PCP-verifiers for NP languages and proving the NP hardness of approximation was established in [10, 2] for the Max-CLIQUE problem (defined below). It seemed at first like an isolated example. The great impact of Sudan et. al.'s [1] theorem was in showing this was not the case. They showed that proving characterization of NP as $PCP(logn, 1)$ implies the NP hardness of approximation for a collection of NP-complete problems including Max-3-SAT, Max-VERTEX COVER, and others (as well as improving the Max-CLIQUE hardness factor).

The basic idea is the following: A PCP type theorem provides a natural(?) optimization problem which cannot be efficiently approximated by any factor better than 2 as follows. Fix a PCP-verifier V_L for an NP language L and an x. Any candidate PCP-witness O_x for x defines an acceptance probability of $V_L^{O_x}(x)$. The gap of $1/2$ in the maximum acceptance probability for $x \in L$ versus $x \notin L$ (which exists by the definition of PCP) implies that it is NP-hard to 2-approximate the maximum acceptance probability of V_L. In other words, the existence of a polynomial time algorithm to 2-approximate the acceptance probability of x by V_L would imply that $NP = P$.

For different optimization problems, showing hardness of approximation is done by demonstrating *reductions* from variants of the above optimization problem. These reductions are far more complex than reductions showing NP-hardness for exact problems as one needs to address the difference in in-approximability factors of problems being reduced to each other.

Moreover, these new NP-hardness results have brought on a surge of new research in the algorithmic community as well. New approximation algorithms have been designed which at times have risen to the task of meeting from above the

approximation factors which were proved using PCP theorems to be best possible (unless $NP = P$) . This has brought on a meeting of two communities of researchers: the algorithm designers and complexity theorists. The former may take the failure of the latter to prove NP hardness of approximating a problem within a particular approximation factor as indication of what factor is feasible and vice versa.

This radical advance is best illustrated by way of a few examples. Finding the exact optimal solution to all of the following problems is $\mathcal{NP}$-complete. Naive approximation algorithms existed for a long time, which no one could improve. They yield completely different approximation factors. For some of these problems we now have essentially found optimal approximation problem. Any further advancement will imply that NP problems are efficiently solvable.

Max-CLIQUE: Given a finite graph on n vertices, find the size of the largest complete subgraph. A single vertex solution is within factor n of optimal. More elaborate algorithms give factor $n^{.999}$. This problem was the first one to be proved hard to approximate using PCP type theorem [10]. It is now known that achieving a factor of $n^{1-\epsilon}$ is $\mathcal{NP}$-hard for every $\epsilon > 0$ [19].

Max-3-SAT: Given a logical formula in conjunctive normal form with n variables where there is at most 3 literals per clause, determine the maximal number of clauses which can be satisfied simultaneously by a single truth assignment. A simple probabilistic algorithm satisfies $\frac{1}{2}$ of the clauses. It is now known [20] that achieving a factor $7/8 - \epsilon$ for $\epsilon > 0$ approximation factor is NP-hard even if the formula is satisfiable. At the same time [21] has shown an algorithm which matches the $7/8$ approximation factor when the formula is satisfiable.

Min-Set Cover: Given a collection of subsets of a given finite universe of size n, determine the size of the smallest subcollection that covers every element in the universe. A simply greedy algorithm, choosing the subsets which maximizes the coverage of as many yet uncovered elements as possible, yields a factor $\ln n$ from optimal. It is now known that approximation by a factor of $(1-\epsilon)\ln n$ is $\mathcal{NP}$-hard for every $\epsilon > 0$ [9].

We point the reader to a collection of papers and expositions by Sudan himself [31] on these works as well as exciting further developments.

References

[1] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, May 1998.

[2] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of = NP. *Journal of the ACM*, 45(1):70–122, January 1998.

[3] L. Babai, and S. Moran, *Arthur-Merlin Games: a = randomized proof system, and a hierarchy of complexity classes*, to appear in J. of Computer Science and Systems. Previous version entitled *Trading Group Theory for Randomness*, Proc. 17th ACM Symp. on Theory of Computing (1985) 421-429.

[4] G. Brassard and C. Creapau, *Non-Transitive Transfer of Confidence: A Perfect*

Zero-Knowledge Interactive Protocol for SAT and Beyond, Proc. of 27th IEEE Symp. on Foundations of Computer Science, (1986).

[5] L. Babai, L. Fortnow, and C. Lund. Non-deterministic exponential time has two-prover interactive protocols. *Computational Complexity*, 1:3–40, 1991.

[6] L. Babai, L. Fortnow, L. Levin, and M. Szegedy. Checking computations in poly-logarithmic time. In *Proc. 23rd ACM Symp. on Theory of Computing*, 1991.

[7] M. Ben-or, S. Goldwasser, J. Kilian, and A. Wigderson. Multi prover interactive proofs: How to remove intractability. In *Proc. 20th ACM Symp. on Theory of Computing*, pages 113–131, 1988.

[8] S. Cook, *The Complexity of Theorem-Proving Procedures* Proc of 3rd Symposium of Theory of Computation, (1971)pp. 151-158.

[9] Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

[10] Uriel Feige, Shafi Goldwasser, Laszlo Lovasz, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM*, 43(2):268–292, 1996.

[11] Katalin Friedl and Madhu Sudan.

[12] ome improvements to total degree tests.

[13] roceedings of the 3rd Annual Israel Symposium on Theory of Computing and Systems, pages 190-198, Tel Aviv, Israel, 4-6 January 1995.

[14] L. Fortnow, J. Rompel, and M. Sipser. On the power of multi-prover interactive protocols. In *Proc. 3rd STRUCTURES*, pages 156–161, 1988.

[15] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, 1979.

[16] S. Goldwasser, S. Micali, and C. Rackoff, *The Knowledge Complexity of Interactive Proof Systems*, Proc. 27th Foundation of Computer Science Conf. (1985), pp. 291-304. Earlier Version: *Knowledge Complexity*,

[17] Oded Goldreich and Madhu Sudan. Locally testable codes and PCPs of almost-linear length. Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science, (to appear,) Vancouver, Canada, 16-19 November, 2002.

[18] Oded Goldreich, Howard J. Karloff, Leonard J. Schulman, Luca Trevisan. Lower Bounds for Linear Locally Decodable Codes and Private Information Retrieval. IEEE Conference on Computational Complexity 2002: 175-183

[19] Johan Håstad. Clique is hard to approximate within n to the power 1-epsilon. *Acta Mathematica*, 182:105–142, 1999.

[20] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48:798–859, 2001.

[21] Howard Karloff, Uri Zwick, *A 7/8-approximation algorithm for MAX 3SAT?* Proc. of 38th FOCS (1997), 406-415.

[22] Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[23] J. Katz and L. Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In Proc. of 32nd ACM STOC, 2000.
[24] C. Lund, L. Fortnow, H. Karloff, and N. Nisan. Algebraic methods for interactive proof systems. *JACM*, 39 (1992), 859–868.
[25] L. A. Levin. Universal sequential search problems. Problems of Information Transmission, 9(3):265-266, 1973.
[26] Silvio Micali. CS proofs. In Proceedings of 35th FOCS, pages 436-453. IEEE Computer Society Press, 1994.
[27] C. Papadimitriou and M Yannakakis. Optimization, approximation and complexity classes. *J. Computer and System Sci.* 43 (1991), 425–440. 1988.
[28] Madhu Sudan. Probabilistically checkable proofs. Scribed notes by Venkatesan Guruswami. To appear in Lecture Notes of IAS/Park City Summer School on Complexity Theory.
[29] Madhu Sudan. Decoding of Reed-Solomon codes beyond the error-correction bound. *Journal of Complexity*, 13(1):180–193, 1997.
[30] Madhu Sudan. List decoding: Algorithms and applications. *SIGACT News*, 31:16–27, 2000.
[31] Madhu Sudan. `theory.lcs.mit.edu/~madhu.`
[32] A. Shamir. IPPSPACE. *JACM*, 39 (1992), 869–877.

Madhu Sudan

Associate Professor of Electrical Engineering and Computer Science, MIT

1987	B.Tech. in Computer Science, Indian Institute of Technology at New Delhi
1992	Ph.D. in Computer Science, University of California at Berkeley
1992–1997	Research Staff Member, IBM Thomas J. Watson Research Center Mathematical Sciences Department
1997	Associate Professor, Department of Electrical Engineering and Computer Science at Massachussetts Institute of Technology

Areas of Special Interests

Theoretical Computer Science, Algorithms, Computational Complexity, Optimization, Coding Theory.

M. Sudan (left) and S. Goldwasser

Invited One-Hour Plenary Lectures

ICM 2002 · Vol. I · 119–135

Discrete Mathematics: Methods and Challenges

Noga Alon*

Abstract

Combinatorics is a fundamental mathematical discipline as well as an essential component of many mathematical areas, and its study has experienced an impressive growth in recent years. One of the main reasons for this growth is the tight connection between Discrete Mathematics and Theoretical Computer Science, and the rapid development of the latter. While in the past many of the basic combinatorial results were obtained mainly by ingenuity and detailed reasoning, the modern theory has grown out of this early stage, and often relies on deep, well developed tools. This is a survey of two of the main general techniques that played a crucial role in the development of modern combinatorics; algebraic methods and probabilistic methods. Both will be illustrated by examples, focusing on the basic ideas and the connection to other areas.

2000 Mathematics Subject Classification: 05-02.
Keywords and Phrases: Combinatorial nullstellensatz, Shannon capacity, Additive number theory, List coloring, The probabilistic method, Ramsey theory, Extremal graph theory.

1. Introduction

The originators of the basic concepts of Discrete Mathematics, the mathematics of finite structures, were the Hindus, who knew the formulas for the number of permutations of a set of n elements, and for the number of subsets of cardinality k in a set of n elements, already in the sixth century. The beginning of Combinatorics as we know it today started with the work of Pascal and De Moivre in the 17th century, and continued in the 18th century with the seminal ideas of Euler in Graph Theory, with his work on partitions and their enumeration, and with his interest in latin squares. These old results are among the roots of the study of formal methods of enumeration, the development of configurations and designs, and the extensive

*School of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel 69978. E-mail: noga@math.tau.ac.il

work on Graph Theory in the last two centuries. The tight connection between Discrete Mathematics and Theoretical Computer Science, and the rapid development of the latter in recent years, led to an increased interest in combinatorial techniques and to an impressive development of the subject. It also stimulated the study of algorithmic combinatorics and combinatorial optimization.

While many of the basic combinatorial results were obtained mainly by ingenuity and detailed reasoning, without relying on many deep, well developed tools, the modern theory has already grown out of this early stage. There are already well developed enumeration methods, some of which are based on deep algebraic tools. The probabilistic method initiated by Erdős (and to some extent, by Shannon) became one of the most powerful tools in the modern theory, and its study has been fruitful to Combinatorics, as well as to Probability Theory. Algebraic and topological techniques play a crucial role in the modern theory, and Polyhedral Combinatorics, Linear Programming and constructions of designs have been developed extensively. Most of the new significant results obtained in the area are inevitably based on the knowledge of these well developed concepts and techniques, and while there is, of course, still a lot of room for pure ingenuity in Discrete Mathematics, much of the progress is obtained by relying on the fast growing accumulated body of knowledge.

Concepts and questions of Discrete Mathematics appear naturally in many branches of mathematics, and the area has found applications in other disciplines as well. These include applications in Information Theory and Electrical Engineering, in Statistical Physics, in Chemistry and Molecular Biology, and, of course, in Computer Science. Combinatorial topics such as Ramsey Theory, Combinatorial Set Theory, Matroid Theory, Extremal Graph Theory, Combinatorial Geometry and Discrepancy Theory are related to a large part of the mathematical and scientific world, and these topics have already found numerous applications in other fields. A detailed account of the topics, methods and applications of Combinatorics can be found in [35].

This paper is mostly a survey of two of the main general techniques that played a crucial role in the development of modern combinatorics; algebraic methods and probabilistic methods. Both will be illustrated by examples, focusing on the basic ideas and the connection to other areas. The choice of topics and examples described here is inevitably biased, and is not meant to be comprehensive. Yet, it hopefully provides some of the flavor of the techniques, problems and results in the area in a way which may be appealing to researchers, even if their main interest is not Discrete Mathematics.

2. Dimension, geometry and information theory

Various algebraic techniques have been used successfully in tackling problems in Discrete Mathematics over the years. These include tools from Representation Theory applied extensively in enumeration problems, spectral techniques used in the study of highly regular structures, and applications of properties of polynomials and tools from algebraic geometry in the theory of Error Correcting Codes and in the study of problems in Combinatorial Geometry. These techniques have numer-

ous interesting applications. Yet, the most fruitful algebraic technique applied in combinatorics, which is possibly also the simplest one, is the so-called dimension argument. In its simplest form, the method can be described as follows. In order to bound the cardinality of a discrete structure A, one maps its elements to vectors in a linear space, and shows that the set A is mapped to a linearly independent set. It then follows that the cardinality of A is bounded by the dimension of the corresponding linear space. This method is often particularly useful in the solution of extremal problems in which the extremal configuration is not unique. The method is effective in such cases because bases in a vector space can be very different from each other and yet all of them have the same cardinality. Many applications of this basic idea can be found in [13], [14], [37].

2.1. Combinatorial geometry

An early application of the dimension argument appears in [49]. A set of points $A \subset R^n$ is a two-distance set if at most 2 distinct positive distances are determined by the points of A. Let $f(n,2)$ denote the maximum possible size of a two-distance set in R^n. The set of all 0/1 vectors in R^{n+1} with exactly two 1's shows that $f(n,2) \geq n(n+1)/2$, and the authors of [49] proved that $f(n,2) \leq (n+1)(n+4)/2$. The upper bound is proved by associating each point of a two-distance set A with a polynomial in n variables, and by showing that these polynomials are linearly independent and all lie in a space of dimension $(n+1)(n+4)/2$. This has been improved by Blokhuis to $(n+1)(n+2)/2$, by showing that one can add $n+1$ additional polynomials that lie in this space to those obtained from the two-distance set, keeping the augmented set linearly independent. See [14] and its references for more details. The precise value of $f(n,2)$ is not known.

Borsuk [21] asked if any compact set of at least 2 points in R^d can be partitioned into at most $d+1$ subsets of smaller diameter. Let $m(d)$ be the smallest integer m so that any such set can be partitioned into a most m subsets of smaller diameter. Borsuk's question is whether $m(d) = d+1$ (the $d+1$ points of a simplex show that $m(d)$ is at least $d+1$.) Kahn and Kalai [42] gave an example showing that this is not the case for all sufficiently large d, by applying a theorem of Frankl and Wilson [33]. Improved versions of their construction have been obtained by Nilli in 1994, by Raigorodski in 1997, by Hinrichs in 2001 and by Hinrichs and Richter in 2002. The last two results are based on some properties of the Leech Lattice and give a construction showing that already in dimension $d = 298$, more than $d+1$ subsets may be needed. All the constructions and the proofs of their properties are based on the dimension argument. Here is a brief sketch of one of them.

Let $n = 4p$, where p is an odd prime, and let $\mathcal{F}$ be the set of all vectors $\mathbf{x} = (x_1, \ldots, x_n) \in \{-1,1\}^n$, where $x_1 = 1$ and the number of negative coordinates of $\mathbf{x}$ is even. One first proves the following.

$$\text{If } \mathcal{G} \subset \mathcal{F} \text{ contains no two orthogonal vectors then } |\mathcal{G}| \leq \sum_{i=0}^{p-1} \binom{n-1}{i}. \qquad (1)$$

This is done by associating each member of $\mathcal{G}$ with a multilinear polynomial of degree at most $p-1$ in $n-1$ variables, so that all the obtained polynomials are

linearly independent. Having established (1), define $S = \{\mathbf{x} * \mathbf{x} : \mathbf{x} \in \mathcal{F}\}$, where $\mathcal{F}$ is as above, and $\mathbf{x} * \mathbf{x}$ is the tensor product of $\mathbf{x}$ with itself, i.e., the vector of length n^2, $(x_{ij} : 1 \leq i, j \leq n)$, where $x_{ij} = x_i x_j$. The norm of each vector in S is n and the scalar product between any two members of S is non-negative. Moreover, by (1) any set of more than $\sum_{i=0}^{p-1} \binom{n-1}{i}$ members of S contains an orthogonal pair, i.e., two points the distance between which is the diameter of S. It follows that S cannot be partitioned into less than $2^{n-2} / \sum_{i=0}^{p-1} \binom{n-1}{i}$ subsets of smaller diameter. This shows that $m(d) \geq c_1^{\sqrt{d}}$ for some $c_1 > 1$. An upper bound of $m(d) \leq c_2^d$ where $c_2 = \sqrt{3/2} + o(1)$ is known, but determining the correct order of magnitude of $m(d)$ is an open question. The following conjecture seems plausible.

Conjecture 2.1 *There is a constant $c > 1$ such that $m(d) > c^d$ for all $d \geq 1$.*

An *equilateral set* (or a simplex) in a metric space, is a set A, so that the distance between any pair of distinct members of A is b, where $b \neq 0$ is a constant. Trivially, the maximum cardinality of such a set in R^n with respect to the (usual) l_2-norm is $n+1$. Somewhat surprisingly, the situation is far more complicated for the l_1 norms. The l_1-distance between two points $\vec{a} = (a_1, \ldots a_n)$ and $\vec{b} = (b_1, \ldots, b_n)$ in R^n is $\|\vec{a} - \vec{b}\|_1 = (\sum_{k=1}^{n} |a_i - b_i|$. Let $e(l_1^n)$ denote the maximum possible cardinality of an equilateral set in l_1^n. The set of standard basis vectors and their negatives shows that $e(l_1^n) \geq 2n$. Kusner [39] conjectured that this is tight, i.e., that $e(l_1^n) = 2n$ for all n. For $n \leq 4$ this is proved in [44]. For general n, the best known upper bound is $e(l_1^n) \leq c_1 n \log n$ for some absolute positive constant c_1. This is proved in [9] by an appropriate dimension argument. Each vector in an equilateral set of m vectors in R^n is mapped to a vector in l_2^t for an appropriate $t = t(m, n)$, by applying a probabilistic technique involving randomized rounding. It is then shown, using a simple argument based on the eigenvalues of the Gram matrix of these new vectors, that they span a space of dimension at least $c_2 m$, implying that $c_2 m \leq t(m, n)$ and supplying the desired result. The precise details require some work, and can be found in [9].

2.2. Capacities and graph powers

Let $G = (V, E)$ be a simple, undirected graph. The power G^n of G is the graph whose set of vertices is V^n in which two distinct vertices $(u_1, u_2, \ldots, u_n)$ and $(v_1, v_2, \ldots, v_n)$ are adjacent iff for all i between 1 and n either $u_i = v_i$ or $u_i v_i \in E$. The *Shannon capacity* $c(G)$ of G is the limit $\lim_{n \to \infty} (\alpha(G^n))^{1/n}$, where $\alpha(G^n)$ is the maximum size of an independent set of vertices in G^n. This limit exists, by super-multiplicativity, and it is always at least $\alpha(G)$.

The study of this parameter was introduced by Shannon in [61], motivated by a question in Information Theory. Indeed, if V is the set of all possible letters a channel can transmit in one use, and two letters are adjacent if they may be confused, then $\alpha(G^n)$ is the maximum number of messages that can be transmitted in n uses of the channel with no danger of confusion. Thus $c(G)$ represents the number of distinct messages *per use* the channel can communicate with no error while used many times.

Calculation of $c(G)$ seems to be very hard. For example $c(C_5) = \sqrt{5}$ was only shown in 1979 by Lovász [50], and $c(C_7)$ remains unknown. Certain polynomially computable upper bounds on $c(G)$ are known including Lovász's theta function $\theta(G)$, and other upper bounds are due to Haemers and to Schrijver.

Another upper bound, based on the dimension argument and related to the bound of Haemers [40], is described in [3], where it is applied to solve a problem of Shannon on the capacity of the disjoint union of two graphs. The *(disjoint) union* of two graphs G and H, denoted by $G+H$, is the graph whose vertex set is the disjoint union of the vertex sets of G and of H and whose edge set is the (disjoint) union of the edge sets of G and H. If G and H are graphs of two channels, then their union represents the *sum* of the channels corresponding to the situation where either one of the two channels may be used, a new choice being made for each transmitted letter. Shannon proved that for every G and H, $c(G+H) \geq c(G) + c(H)$ and that equality holds in many cases. He conjectured that in fact equality always holds. In [3] it is shown that this is false in the following strong sense.

Theorem 2.2 *For every k there is a graph G so that the Shannon capacity of the graph and that of its complement $\overline{G}$ satisfy $c(G) \leq k, c(\overline{G}) \leq k$, whereas $c(G+\overline{G}) \geq k^{(1+o(1))\frac{\log k}{8 \log \log k}}$ and the $o(1)$-term tends to zero as k tends to infinity.*

Therefore, the capacity of the disjoint union of two graphs can be much bigger than the capacity of each of the two graphs. Strangely enough, it is not even known if the maximum possible capacity of a disjoint union of two graphs G and H, each of capacity at most k, is bounded by any function of k. It seems very likely that this is the case.

3. Polynomials, addition and graph coloring

The study of algebraic varieties, that is, sets of common roots of systems of polynomials, is the main topic of algebraic geometry. The most elementary property of a univariate nonzero polynomial over a field is the fact that it does not have more roots than its degree. This elementary property is surprisingly effective in Combinatorics: it plays a major role in the theory of error correcting codes, and has many applications in the study of finite geometries — see, e.g., [14]. A similar property holds for polynomials of several variables, and can also be used to supply results in Discrete Mathematics. In this section we describe a general result of this type, which is called in [4] *Combinatorial Nullstellensatz*, and briefly sketch some of its applications in Additive Number Theory and in Graph Theory.

3.1. Combinatorial nullstellensatz

Hilbert's Nullstellensatz (see, e.g., [65]) is the fundamental theorem that asserts that if F is an algebraically closed field, and $f, g_1, \ldots, g_m$ are polynomials in the ring of polynomials $F[x_1, \ldots, x_n]$, where f vanishes over all common zeros of $g_1, \ldots, g_m$, then there is an integer k and polynomials $h_1, \ldots, h_m$ in $F[x_1, \ldots, x_n]$

so that

$$f^k = \sum_{i=1}^{n} h_i g_i.$$

In the special case $m = n$, where each g_i is a univariate polynomial of the form $\prod_{s \in S_i}(x_i - s)$ for some $S_i \subset F$, a stronger conclusion holds. It can be shown that if F is an arbitrary field, f, g_i, S_i are as above, and f vanishes over all the common zeros of $g_1, \ldots, g_n$ (that is; $f(s_1, \ldots, s_n) = 0$ for all $s_i \in S_i$), then there are polynomials $h_1, \ldots, h_n \in F[x_1, \ldots, x_n]$ satisfying $deg(h_i) \leq deg(f) - deg(g_i)$ so that

$$f = \sum_{i=1}^{n} h_i g_i.$$

As a consequence of the above one can prove the following.

Theorem 3.1 *Let F be an arbitrary field, and let $f = f(x_1, \ldots, x_n)$ be a polynomial in $F[x_1, \ldots, x_n]$. Suppose the degree $deg(f)$ of f is $\sum_{i=1}^{n} t_i$, where each t_i is a nonnegative integer, and suppose the coefficient of $\prod_{i=1}^{n} x_i^{t_i}$ in f is nonzero. If $S_1, \ldots, S_n$ are subsets of F with $|S_i| > t_i$, then there are $s_1 \in S_1, s_2 \in S_2, \ldots, s_n \in S_n$ so that*

$$f(s_1, \ldots, s_n) \neq 0.$$

The detailed proof, as well as many applications, can be found in [4]. A quick application, first proved in [5], is the assertion that for any prime p, any loopless graph $G = (V, E)$ with average degree bigger than $2p - 2$ and maximum degree at most $2p - 1$ contains a p-regular subgraph.

To prove it, let $(a_{v,e})_{v \in V, e \in E}$ denote the incidence matrix of G defined by $a_{v,e} = 1$ if $v \in e$ and $a_{v,e} = 0$ otherwise. Associate each edge e of G with a variable x_e and consider the polynomial

$$f = \prod_{v \in V} [1 - (\sum_{e \in E} a_{v,e} x_e)^{p-1}] - \prod_{e \in E} (1 - x_e),$$

over $GF(p)$. Applying Theorem 3.1 with $t_i = 1$ and $S_i = \{0, 1\}$ for all i, we conclude that there are values $x_e \in \{0, 1\}$ such that $f(x_e : e \in E) \neq 0$. It is now easy to check that in the subgraph consisting of all edges $e \in E$ for which $x_e = 1$ all degrees are divisible by p, and since the maximum degree is smaller than $2p$ all positive degrees are precisely p, as needed.

Pyber applied the above result to solve a problem of Erdős and Sauer and prove that any simple graph on n vertices with at least $200n \log n$ edges contains a 3-regular subgraph. Pyber, Rödl and Szemerédi proved that this is not very far from being best possible, by showing, using probabilistic arguments, that there are simple graphs on n vertices with at least $cn \log \log n$ edges that contain no 3-regular subgraphs. See [58] for some further related results.

3.2. Additive number theory

The Cauchy-Davenport Theorem, which has numerous applications in Additive Number Theory, is the statement that if p is a prime, and A, B are two nonempty subsets of Z_p, then

$$|A+B| \geq min\{p, |A|+|B|-1\}.$$

Cauchy proved this theorem in 1813, and applied it to give a new proof to a lemma of Lagrange in his well known 1770 paper that shows that every positive integer is a sum of four squares. Davenport formulated the theorem as a discrete analogue of a conjecture of Khintchine about the Schnirelman density of the sum of two sequences of integers. There are numerous extensions of this result, see, e.g., [56]. A simple algebraic proof of this result is given in [7], and its main advantage is that it extends easily and gives several related results. This proof can be described as a simple application of Theorem 3.1. If $|A|+|B| > p$, then the result is trivial, as the sets A and $g-B$ intersect, for each $g \in Z_p$. Otherwise, assuming the result is false and $|A+B| \leq |A|+|B|-2$, let C be a subset of Z_p satisfying $A+B \subset C$ and $|C| = |A|+|B|-2$. Define $f = f(x,y) = \prod_{c \in C}(x+y-c)$ and apply Theorem 3.1 with $t_1 = |A|-1, t_2 = |B|-1$, $S_1 = A, S_2 = B$ to get a contradiction.

Using similar (though somewhat more complicated) arguments, the following related result is proved in [7].

Proposition 3.2 *Let p be a prime, and let $A_0, A_1, \ldots, A_k$ be nonempty subsets of the cyclic group Z_p. If $|A_i| \neq |A_j|$ for all $0 \leq i < j \leq k$ and $\sum_{i=0}^{k} |A_i| \leq p + \binom{k+2}{2} - 1$ then*

$$|\{a_0 + a_1 + \ldots + a_k : a_i \in A_i, a_i \neq a_j \quad \textit{for all} \quad i \neq j\}| \geq \sum_{i=0}^{k} |A_i| - \binom{k+2}{2} + 1.$$

The very special case of this proposition in which $k=1$, $A_0 = A$ and $A_1 = A - \{a\}$ for an arbitrary element $a \in A$ implies that if $A \subset Z_p$ and $2|A| - 1 \leq p+2$ then the number of sums $a_1 + a_2$ with $a_1, a_2 \in A$ and $a_1 \neq a_2$ is at least $2|A|-3$. This supplies a short proof of a result of Dias Da Silva and Hamidoune [23], which settles a conjecture of Erdős and Heilbronn (cf., e.g., [27]).

Snevily [62] conjectured that for any two sets A and B of equal cardinality in any abelian group of odd order, there is a renumbering a_i, b_i of the elements of A and B so that all sums $a_i + b_i$ are pairwise distinct.

For the cyclic group Z_p of prime order, this follows easily from Theorem 3.1 by considering the polynomial $f = \prod_{i<j}(x_i - x_j) \prod_{i<j}(a_i + x_i - a_j - x_j)$ with $S_1 = \cdots = S_k = B$.

More generally, Dasgupta et al. [24] proved the conjecture for any cyclic group of odd order, by applying the polynomial method for polynomials over $Q[\omega]$, where ω is an appropriate root of unity, and by considering G as a subgroup of the multiplicative group of this field. Further related results appear in [63].

Additional applications of Theorem 3.1 in additive number theory can be found in [4].

3.3. Graph coloring

Theorem 3.1 has various applications in the study of Graph Coloring, which is the most popular area in Graph Theory. We sketch below the basic approach, following [12]. See also [52], [53] for a related method.

A *vertex coloring* of a graph G is an assignment of a color to each vertex of G. The coloring is *proper* if adjacent vertices get distinct colors. The *chromatic number* $\chi(G)$ of G is the minimum number of colors used in a proper vertex coloring of G. An *edge coloring* of G is, similarly, an assignment of a color to each edge of G. It is *proper* if adjacent edges receive distinct colors. The minimum number of colors in a proper edge coloring of G is the *chromatic index* $\chi'(G)$ of G. This is equal to the chromatic number of the line graph of G.

A graph $G = (V, E)$ is *k-choosable* if for every assignment of sets of integers $S(v) \subset Z$, each of size k, to the vertices $v \in V$, there is a proper vertex coloring $c : V \mapsto Z$ so that $c(v) \in S(v)$ for all $v \in V$. The *choice number* of G, denoted by $ch(G)$, is the minimum integer k so that G is k-choosable. Obviously, this number is at least the chromatic number $\chi(G)$ of G. The choice number of the line graph of G, denoted by $ch'(G)$, is usually called the *list chromatic index* of G, and it is clearly at least the chromatic index $\chi'(G)$ of G.

The study of choice numbers was introduced, independently, by Vizing [67] and by Erdős, Rubin and Taylor [29]. There are many graphs G for which the choice number $ch(G)$ is strictly larger than the chromatic number $\chi(G)$ (a complete bipartite graph with 3 vertices in each color class is one such example). In view of this, the following conjecture, suggested independently by various researchers including Vizing, Albertson, Collins, Tucker and Gupta, which apparently appeared first in print in [17], is somewhat surprising.

Conjecture 3.3 (The list coloring conjecture) *For every graph G, $ch'(G) = \chi'(G)$.*

This conjecture asserts that for *line graphs* there is no gap at all between the choice number and the chromatic number. Many of the most interesting results in the area are proofs of special cases of this conjecture, which is still wide open.

The *graph polynomial* $f_G = f_G(x_1, x_2, \ldots, x_n)$ of a graph $G = (V, E)$ on a set $V = \{1, \ldots, n\}$ of n vertices is defined by $f_G(x_1, x_2, \ldots, x_n) = \Pi\{(x_i - x_j) : i < j \ , \ ij \in E\}$. This polynomial has been studied by various researchers, starting already with Petersen [57] in 1891.

Note that if $S_1, \ldots, S_n$ are sets of integers, then there is a proper coloring assigning to each vertex i a color from its list S_i, if and only if there are $s_i \in S_i$ such that $f_G(s_1, \ldots, s_n) \neq 0$. This condition is precisely the one appearing in the conclusion of Theorem 3.1, and it is therefore natural to expect that this theorem can be useful in tackling coloring problems. By applying it to line graphs of planar cubic graphs, and by interpreting the appropriate coefficient of the corresponding polynomial combinatorially, it can be shown, using a known result of Vigneron [66] and the Four Color Theorem, that the list chromatic index of every 2-connected cubic planar graph is 3. This is a strengthening of the Four Color Theorem, which is well known to be equivalent to the fact that the chromatic index of any such graph is 3. An extension of this result appears in [25].

Additional results on graph coloring and choice numbers using the above algebraic approach are described in the survey [2]. These include the fact that the choice number of every planar bipartite graph is at most 3, thus solving a conjecture raised in [29], and the assertion, proved in [32], that if G is a graph on $3n$ vertices, whose set of edges is the disjoint union of a Hamilton cycle and n pairwise vertex-disjoint triangles, then the choice number and the chromatic number of G are both 3.

4. The probabilistic method

The discovery that deterministic statements can be proved by probabilistic reasoning, led already in the middle of the previous century to several striking results in Analysis, Number Theory, Combinatorics and Information Theory. It soon became clear that the method, which is now called *the probabilistic method*, is a very powerful tool for proving results in Discrete Mathematics. The early results combined combinatorial arguments with fairly elementary probabilistic techniques, whereas the development of the method in recent years required the application of more sophisticated tools from Probability Theory. In this section we illustrate the method and describe several recent results. More material can be found in the recent books [11], [16], [41] and [55].

4.1. Thresholds for random properties

The systematic study of Random Graphs was initiated by Erdős and Rényi whose first main paper on the subject is [28]. Formally, $G(n,p)$ denotes the probability space whose points are graphs on a fixed set of n labelled vertices, where each pair of vertices forms an edge, randomly and independently, with probability p. The term "the random graph $G(n,p)$" means, in this context, a random point chosen in this probability space. Each graph property A (that is, a family of graphs closed under graph isomorphism) is an event in this probability space, and one may study its probability $Pr[A]$, that is, the probability that the random graph $G(n,p)$ lies in this family. In particular, we say that A holds *almost surely* if the probability that $G(n,p)$ satisfies A tends to 1 as n tends to infinity. There are numerous papers dealing with random graphs, and the two recent books [16], [41] provide excellent extensive accounts of the known results in the subject.

One of the important discoveries of Erdös and Rényi was the discovery of *threshold functions*. A function $r(n)$ is called a threshold function for a graph property A, if when $p(n)/r(n)$ tends to 0, then $G(n,p(n))$ does not satisfy A almost surely, whereas when $p(n)/r(n)$ tends to infinity, then $G(n,p(n))$ satisfies A almost surely. Thus, for example, they identified the threshold function for the property of being connected very precisely: if $p(n) = \frac{\ln n}{n} + \frac{c}{n}$, then, as n tends to infinity, the probability that $G(n,p(n))$ is connected tends to $e^{-e^{-c}}$.

A graph property is *monotone* if it is closed under the addition of edges. Note that many interesting graph properties, like hamiltonicity, non-planarity, connectivity or containing at least 10 vertex disjoint triangles are monotone.

Bollobás and Thomason [18] proved that *any* monotone graph property has a threshold function. Their proof applies to any monotone family of subsets of a finite set, and holds even without the assumption that the property A is closed under graph isomorphism.

Friedgut and Kalai [30] showed that the symmetry of graph properties can be applied to obtain a sharper result. They proved that for any monotone graph property A, if $G(n,p)$ satisfies A with probability at least ϵ, then $G(n,q)$ satisfies A with probability at least $1-\epsilon$, for $q = p + O(\log(1/2\epsilon)/\log n)$.

The proof follows by combining two results. The first is a simple but fundamental lemma of Margulis [51] and Russo [60], which is useful in Percolation Theory. This lemma can be used to express the derivative with respect to p of the probability that $G(n,p)$ satisfies A as a sum of contributions associated with the single potential edges. The second result is a theorem of [19], which is proved using Harmonic Analysis, that asserts that at least one such contribution is always large. The symmetry implies that all contributions are the same and the result follows. See also [64] for some related results. These results hold for every transitive group of symmetries. In [20] it is shown that one can, in fact, prove that the threshold for graph properties is even sharper, by taking into account the precise group of symmetries induced on the edges of the complete graph by permuting the vertices. It turns out that for every monotone graph property and for every fixed $\epsilon > 0$, the width of the interval in which the probability the property holds increases from ϵ to $1-\epsilon$ is at most $c_\delta/(\log n)^{2-\delta}$ for all $\delta > 0$. The power 2 here is tight, as shown by the property of containing a clique of size, say, $\lfloor 2\log_2 n \rfloor$.

It is natural to call the threshold for a monotone graph property *sharp* if for every fixed positive ϵ, the width w of the interval in which the probability that the property holds increases from ϵ to $1-\epsilon$ satisfies $w = o(p)$, where p is any point inside this interval. In [31] Friedgut obtained a beautiful characterization of all monotone graph properties for which the threshold is sharp. Roughly speaking, a property does not have a sharp threshold if and only if it can be approximated well in the relevant range of the probability p by a property that is determined by constant size witnesses. Thus, for example, the property of containing 5 vertex disjoint triangles does not have a sharp threshold, whereas the property of having chromatic number bigger than 10 does. A similar result holds for hypergraphs as well. The proofs combine probabilistic and combinatorial arguments with techniques from Harmonic analysis.

4.2. Ramsey numbers

Let $H_1, H_2, \ldots, H_k$ be a sequence of k finite, undirected, simple graphs. The (multicolored) *Ramsey number* $r(H_1, H_2, \ldots, H_k)$ is the minimum integer r such that in every edge coloring of the complete graph on r vertices by k colors, there is a monochromatic copy of H_i in color i for some $1 \leq i \leq k$. By a (special case of) a well known theorem of Ramsey (c.f., e.g., [38]), this number is finite for every sequence of graphs H_i.

The determination or estimation of these numbers is usually a very difficult problem. When all graphs H_i are complete graphs with more than two vertices, the

only values that are known precisely are those of $r(K_3, K_m)$ for $m \leq 9$, $r(K_4, K_4)$, $r(K_4, K_5)$ and $r(K_3, K_3, K_3)$. Even the determination of the asymptotic behaviour of Ramsey numbers up to a constant factor is a hard problem, and despite a lot of efforts by various researchers (see, e.g., [38], [22] and their references), there are only a few infinite families of graphs for which this behaviour is known.

In one of the first applications of the probabilistic method in Combinatorics, Erdős [26] proved that if $\binom{n}{k}2^{1-\binom{k}{2}} < 1$ then $R(k,k) > n$, that is, there exists a 2-coloring of the edges of the complete graph on n vertices containing no monochromatic clique of size k. The proof is extremely simple; the probability that a random two-edge coloring of K_n contains a monochromatic K_k is at most $\binom{n}{k}2^{1-\binom{k}{2}} < 1$, and hence there is a coloring with the required property.

A particularly interesting example of an infinite family for which the asymtotic behaviour of the Ramsey number is known, is the following result of Kim [43] together with that of Ajtai, Komlós and Szemerédi [1].

Theorem 4.1 ([43], [1]) *There are two absolute positive constants c_1, c_2 such that*

$$c_1 m^2 / \log m \leq r(K_3, K_m) \leq c_2 m^2 / \log m$$

for all $m > 1$.

The upper bound, proved in [1], is probabilistic, and applies a certain random greedy algorithm. The lower bound is proved by a "semi-random" construction and proceeds in stages. The detailed analysis is subtle, and is based on certain large deviation inequalities.

Even less is known about the asymptotic behaviour of multicolored Ramsey numbers, that is, Ramsey numbers with at least 3 colors. The asymptotic behaviour of $r(K_3, K_3, K_m)$, for example, has been very poorly understood until recently, and Erdős and Sós conjectured in 1979 (c.f., e.g., [22]) that

$$\lim_{m \mapsto \infty} \frac{r(K_3, K_3, K_m)}{r(K_3, K_m)} = \infty.$$

This has been proved recently, in a strong sense, in [10], where it is shown that in fact $r(K_3, K_3, K_m)$ is equal, up to logarithmic factors, to m^3. A more complicated, related result proved in [10], that supplies the asymptotic behaviour of infinitely many families of Ramsey numbers up to a constant factor is the following.

Theorem 4.2 *For every $t > 1$ and $s \geq (t-1)! + 1$ there are two positive constants c_1, c_2 such that for every $m > 1$*

$$c_1 \frac{m^t}{\log^t m} \leq r(K_{t,s}, K_{t,s}, K_{t,s}, K_m) \leq c_2 \frac{m^t}{\log^t m},$$

where $K_{t,s}$ is the complete bipartite graph with t vertices in one color class and s vertices in the other.

The proof combines spectral techniques, character sum estimates, and probabilistic arguments.

4.3. Turán type results

For a graph H and an integer n, the Turán number $ex(n, H)$ is the maximum possible number of edges in a simple graph on n vertices that contains no copy of H. The asymptotic behavior of these numbers for graphs of chromatic number at least 3 is well known, see, e.g., [15]. For bipartite graphs H, however, much less is known, and there are relatively few nontrivial bipartite graphs H for which the order of magnitude of $ex(n, H)$ is known.

A result of Füredi [34] implies that for every fixed bipartite graph H in which the degrees of all vertices in one color class are at most r, there is some $c = c(H) > 0$ such that $ex(n, H) \leq cn^{2-1/r}$. As observed in [6], this result can be derived from a simple and yet surprisingly powerful probabilistic lemma, variants of which have been proved and applied by various researchers starting with Rödl and including Kostochka, Gowers and Sudakov (see [46], [36], [47]). The lemma asserts, roughly, that every graph with sufficiently many edges contains a large subset A in which every a vertices have many common neighbors. The proof uses a process that may be called a *dependent random choice* for finding the set A; A is simply the set of all common neighbors of an appropriately chosen random set R. Intuitively, it is clear that if some a vertices have only a few common neighbors, it is unlikely all the members of R will be chosen among these neighbors. Hence, we do not expect A to contain any such subset of a vertices. This simple idea can be extended. In particular, it can be used to bound the Turán numbers of degenerate bipartite graphs.

A graph is *r-degenerate* if every subgraph of it contains a vertex of degree at most r. An old conjecture of Erdős asserts that for every fixed r-degenerate bipartite graph H, $ex(n, H) \leq O(n^{2-1/r})$, and the above technique suffices to show that there is an absolute constant $c > 0$, such that for every such H, $ex(n, H) \leq n^{2-c/r}$.

Further questions and results about Turán numbers can be found in [6], [15] and their references.

5. Algorithms and explicit constructions

The rapid development of Theoretical Computer Science and its tight connection to Discrete Mathematics motivated the study of the algorithmic aspects of algebraic and probabilistic techniques. Can a combinatorial structure, or a substructure of a given one, whose existence is proved by algebraic or probabilistic means, be constructed *explicitly* (that is, by an efficient deterministic algorithm)? Can the algorithmic problems corresponding to existence proofs be solved by efficient procedures? The study of these questions often requires tools from other branches of mathematics.

As described in subsection 3.3, if G is a graph on $3n$ vertices, whose set of edges is the disjoint union of a Hamilton cycle and n pairwise vertex-disjoint triangles, then the chromatic number of G is 3. Can we solve the corresponding algorithmic problem efficiently ? That is, is there a polynomial time, deterministic or randomized algorithm, that given an input graph as above, colors it properly with 3 colors? Similarly, as mentioned in subsection 3.3, the list chromatic index of any planar

cubic 2-connected graph is 3. Can we color properly the edges of any given planar cubic 2-connected graph using given lists of three colors per edge, in polynomial time?

These problems, as well as the algorithmic versions of additional applications of Theorem 3.1, are open. Of course, an algorithmic version of the theorem itself would provide efficient procedures for solving all these questions. The input for such an algorithm is a polynomial in n variables over a field described, say, by a polynomial size arithmetic circuit. Suppose that this polynomial satisfies the assumptions of Theorem 3.1, and that the fact it satisfies it can be checked efficiently. The algorithm should then find, efficiently, a point $(s_1, s_2, \ldots, s_n)$ satisfying the conclusion of Theorem 3.1.

Unfortunately, it seems unlikely that such a general result can exist, as it would imply that there are no one-way permutations. Indeed, let $F : \{0,1\}^n \mapsto \{0,1\}^n$ be a $1-1$ function, and suppose that for any $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, the value of $F(x)$ can be computed efficiently. Every Boolean function can be expressed as a multilinear polynomial over $GF(2)$, and hence, when we wish to find an x such that $F(x) = y = (y_1, \ldots, y_n)$, we can write it as a system of multilinear polynomials over $GF(2)$: $F_i(x) = y_i$ for all $1 \leq i \leq n$. Equivalently, this can be written as $\prod_{i=1}^{n}(F_i(x) + y_i + 1) \neq 0$. This last equation has a unique solution, implying that its left hand side, written as a multilinear polynomial, is of full degree n (since otherwise it is easy to check that it attains the value 1 an even number of times). It follows that the assumptions of Theorem 3.1 with $f = \prod_{i=1}^{n}(F_i(x) + y_i + 1)$, $t_i = 1$ and $S_i = GF(2)$ hold. Thus, the existence of an efficient algorithm as above would enable us to invert F efficiently, implying that there cannot be any one-way permutations. As this seems unlikely, it may be more productive (and yet challenging) to try and develop efficient procedures for solving the particular algorithmic problems corresponding to the results obtained by the theorem.

Probabilistic proofs also suggest the study of the corresponding algorithmic problems. This is related to the study of randomized algorithms, a topic which has been developed tremendously during the last decade. See, e.g., [54] and its many references. In particular, it is interesting to find explicit constructions of combinatorial structures whose existence is proved by probabilistic arguments. "Explicit" here means that there is a an efficient algorithm that constructs the desired structure in time polynomial in its size. Constructions of this type, besides being interesting in their own, have applications in other areas. Thus, for example, explicit constructions of error correcting codes that are as good as the random ones are of interest in information theory, and explicit constructions of certain Ramsey type colorings may have applications in derandomization — the process of converting randomized algorithms into deterministic ones.

It turns out, however, that the problem of finding a good explicit construction is often very difficult. Even the simple proof of Erdős, described in subsection 4.2, that there are two-edge colorings of the complete graph on $\lfloor 2^{m/2} \rfloor$ vertices containing no monochromatic clique of size m, leads to an open problem which seems very difficult. Can we construct, explicitly, such a coloring of a complete graph on $n \geq (1+\epsilon)^m$ vertices, in time which is polynomial in n, where $\epsilon > 0$ is

any positive absolute constant ?

This problem is still open, despite a lot of efforts. The best known explicit construction is due to Frankl and Wilson [33], who gave an explicit two-edge coloring of the complete graph on $m^{(1+o(1))\frac{\log m}{4 \log \log m}}$ vertices with no monochromatic clique on m vertices.

The construction of explicit two-edge colorings of large complete graphs K_n with no red K_s and no blue K_m for fixed s and large m also appears to be very difficult. Using probabilistic arguments it can be shown that there are such colorings for n which is $c\left(\frac{m}{\log m}\right)^{(s+1)/2}$ for some absolute constant $c > 0$. The best known explicit construction, however, given in [8], works only for $m^{\delta\sqrt{\log s/\log\log s}}$, for some absolute constant $\delta > 0$. The description of the construction is not complicated but the proof of its properties relies on tools from various mathematical areas. These include some ideas from algebraic geometry obtained in [45], the well known bound of Weil on character sums, spectral techniques and their connection to the pseudo-random properties of graphs, the known bounds of [48] for the problem of Zarankiewicz and the well known Erdős-Rado bound for the existence of Δ-systems.

The above example is typical, and illustrates the fact that tools from various mathematical disciplines often appear in the design of explicit constructions of combinatorial structures. Other examples that demonstrate this fact are the construction of Algebraic Geometry codes, and the construction of sparse pseudo-random graphs called expanders.

6. Some future challenges

Several specific open problems in Discrete Mathematics are mentioned throughout this article. These, and many additional ones, provide interesting challenges for future research in the area. We conclude with some brief comments on two more general future challenges.

It seems safe to predict that in the future there will be additional incorporation of methods from other mathematical areas in Combinatorics. However, such methods often provide non-constructive proof techniques, and the conversion of these to algorithmic ones may well be one of the main future challenges of the area. Another interesting recent development is the increased appearance of Computer aided proofs in Combinatorics, starting with the proof of the Four Color Theorem, and including automatic methods for the discovery and proof of hypergeometric identities — see [59]. A successful incorporation of such proofs in the area, without losing its special beauty and appeal, is another challenge. These challenges, the fundamental nature of the area, its tight connection to other disciplines, and the many fascinating specific open problems studied in it, ensure that Discrete Mathematics will keep playing an essential role in the general development of science in the future as well.

References

[1] M. Ajtai, J. Komlós and E. Szemerédi, *A note on Ramsey numbers*, J. Combinatorial Theory Ser. A 29 (1980), 354–360.

[2] N. Alon, *Restricted colorings of graphs*, in *Surveys in Combinatorics*, Proc. 14^{th} British Combinatorial Conference, London Mathematical Society Lecture Notes Series 187, edited by K. Walker, Cambridge University Press, 1993, 1–33.

[3] N. Alon, *The Shannon Capacity of a union*, Combinatorica 18 (1998), 301–310.

[4] N. Alon, *Combinatorial Nullstellensatz*, Combinatorics, Probability and Computing 8 (1999), 7–29.

[5] N. Alon, S. Friedland and G. Kalai, *Regular subgraphs of almost regular graphs*, J. Combinatorial Theory Ser. B 37 (1984), 79–91.

[6] N. Alon, M. Krivelevich and B. Sudakov, *Turán numbers of bipartite graphs and related Ramsey-type questions*, Geom. Funct. analysis, to appear.

[7] N. Alon, M. B. Nathanson and I. Z. Ruzsa, *The polynomial method and restricted sums of congruence classes*, J. Number Theory 56 (1996), 404–417.

[8] N. Alon and P. Pudlak, *Constructive lower bounds for off-diagonal Ramsey numbers*, Israel J. Math. 122 (2001), 243–251.

[9] N. Alon and P. Pudlak, *Equilateral sets in* l_p^n, Geom. Funct. Analysis, to appear.

[10] N. Alon and V. Rödl, *Asymptotically tight bounds for some multicolored Ramsey numbers*, to appear.

[11] N. Alon and J. H. Spencer, **The Probabilistic Method**, Second Edition, Wiley, New York, 2000.

[12] N. Alon and M. Tarsi, *Colorings and orientations of graphs*, Combinatorica 12 (1992), 125–134.

[13] L. Babai and P. Frankl, **Linear Algebra Methods in Combinatorics**, to appear.

[14] A. Blokhuis, *Polynomials in Finite Geometries and Combinatorics*, in *Surveys in Combinatorics*, Proc. 14^{th} British Combinatorial Conference, London Mathematical Society Lecture Notes Series 187, edited by K. Walker, Cambridge University Press, 1993, 35–52.

[15] B. Bollobás, **Extremal Graph Theory**, Academic Press, London, 1978.

[16] B. Bollobás, **Random Graphs**, Second Edition, Academic Press, London, 2001.

[17] B. Bollobás and A. J. Harris, *List colorings of graphs*, Graphs and Combinatorics 1 (1985), 115–127.

[18] B. Bollobás and A. Thomason, *Threshold functions*, Combinatorica 7 (1987), 35–38.

[19] J. Bourgain, J. Kahn, G. Kalai, Y. Katznelson and N. Linial, *The influence of variables in product spaces*, Israel J. Math. 77 (1992), 55–64.

[20] J. Bourgain and G. Kalai, *The influence of variables in product spaces under group symmetries*, GAFA 7 (1997), 438–461.

[21] K. Borsuk, *Drei Sätze über die n-dimensionale euklidische Sphäre*, Fundamenta Math. 20 (1933), 177–190.

[22] F. Chung and R. L. Graham, **Erdős on Graphs: His Legacy of Unsolved Problems**, A. K. Peters, Wellesley, MA, 1998.

[23] J. A. Dias da Silva and Y. O. Hamidoune, *Cyclic spaces for Grassmann derivatives and additive theory*, Bull. London Math. Soc. 26 (1994), 140–146.
[24] S. Dasgupta, G. Károlyi, O. Serra and B. Szegedy, *Transversals of additive Latin squares*, Israel J. Math. 126 (2001), 17–28.
[25] M. N. Ellingham and L. Goddyn, *List edge colourings of some* 1*-factorable multigraphs*, Combinatorica 16 (1996), 343–352.
[26] P. Erdős, *Some remarks on the theory of graphs*, Bulletin of the Amer. Math. Soc. 53 (1947), 292–294.
[27] P. Erdős and R. L. Graham, **Old and New Problems and Results in Combinatorial Number Theory**, L'Enseignement Mathématique, Geneva, 1980.
[28] P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hungar. Acad. Sci.* 5 (1960), 17–61.
[29] P. Erdős, A. L. Rubin and H. Taylor, *Choosability in graphs*, Proc. West Coast Conf. on Combinatorics, Graph Theory and Computing, Congressus Numerantium XXVI, 1979, 125–157.
[30] E. Friedgut and G. Kalai, *Every monotone graph property has a sharp threshold*, Proc. AMS 124 (1996), 2993–3002.
[31] E. Friedgut, *Sharp thresholds of graph properties and the* k*-sat problem* (with an appendix by J. Bourgain), J. Amer. Math. Soc. 12 (1999), 1017–1054.
[32] H. Fleischner and M. Stiebitz, *A solution to a coloring problem of P. Erdős*, Discrete Math. 101 (1992), 39–48.
[33] P. Frankl and R. Wilson, *Intersection theorems with geometric consequences*, Combinatorica 1 (1981), 259–286.
[34] Z. Füredi, *On a Turán type problem of Erdős*, Combinatorica 11 (1991), 75–79.
[35] R. L. Graham, M. Grötschel and L. Lovász, Editors, **Handbook of Combinatorics**, North Holland, Amsterdam, 1995.
[36] W. T. Gowers, *A new proof of Szemerédi's theorem for arithmetic progressions of length four*, Geom. Funct. Analysis 8 (1998), 529–551.
[37] C. Godsil and G. Royle, **Algebraic Graph Theory**, Springer-Verlag, New York, 2001.
[38] R. L. Graham, B. L. Rothschild and J. H. Spencer, **Ramsey Theory**, Second Edition, Wiley, New York, 1990.
[39] R. Guy, editor, *Unsolved Problems: an olla-podrida of open problems, often oddly posed*, Amer. Math. Monthly 90 (1983), 196–200.
[40] W. Haemers, *On some problems of Lovász concerning the Shannon capacity of a graph*, IEEE Trans. Inform. Theory **25** (1979), 231–232.
[41] S. Janson, T. Łuczak and A. Ruciński, **Random Graphs**, Wiley, New York, 2000.
[42] J. Kahn and G. Kalai, *A counterexample to Borsuk's conjecture*, Bulletin of the AMS 29 (1993), 60–62.
[43] J. H. Kim, *The Ramsey number* $R(3,t)$ *has order of magnitude* $t^2/\log t$, Random Structures and Algorithms 7 (1995), 173–207.
[44] J. Koolen, M. Laurent and A. Schrijver, *Equilateral dimension of the rectilinear space*, Designs, Codes and Crypt. 21 (2000), 149–164.

[45] J. Kollár, L. Rónyai and T. Szabó, *Norm-graphs and bipartite Turán numbers*, Combinatorica 16 (1996), 399–406.
[46] A. Kostochka and V. Rödl, *On graphs with small Ramsey numbers*, J. Graph Theory 37 (2001), 198–204.
[47] A. Kostochka and B. Sudakov, *On Ramsey numbers of sparse graphs*, to appear.
[48] T. Kövari, V.T. Sós and P. Turán, *On a problem of K. Zarankiewicz*, Colloquium Math., 3 (1954), 50–57.
[49] D. G. Larman, C. A. Rogers and J. J. Seidel, *On two-distance sets in Euclidean space*, Bull. London Math. Soc. 9 (1977), 261–267.
[50] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory **25** (1979), 1–7.
[51] G. A. Margulis, *Probabilistic characteristics of graphs with large connectivity*, Prob. Peredachi Inform. 10 (1974), 101–108.
[52] Y. Matiyasevich, *A criterion for colorability of vertices of a graph stated in terms of edge orientations* (in Russian), Discrete Analysis (Novosibirsk) 26 (1974), 65–71.
[53] Y. Matiyasevich, *Some algebraic methods for calculation of the number of colorings of a graph*, to appear.
[54] R. Motwani and P. Raghavan, **Randomized Algorithms**, Cambridge University Press, New York, 1995.
[55] M. Molloy and B. Reed, **Graph Coloring and the Probabilistic Method**, Springer-Verlag, Berlin, 2001.
[56] M. B. Nathanson, **Additive Number Theory: Inverse Theorems and the Geometry of Sumsets**, Springer-Verlag, New York, 1996.
[57] J. Petersen, *Die Theorie der regulären Graphs*, Acta Math. 15 (1891), 193–220.
[58] L. Pyber, V. Rödl and E. Szemerédi, *Dense Graphs without 3-regular Subgraphs*, J. Combinatorial Theory Ser. B 63 (1995), 41–54.
[59] M. Petkovsek, H. Wilf and D. Zeilberger, **A=B**, A. K. Peters, Wellesley, 1996.
[60] L. Russo, *On the critical percolation probabilities*, Z. Wahrsch. werw. Gebiete 43 (1978), 39–48.
[61] C. E. Shannon, *The zero-error capacity of a noisy channel*, IRE Trans. Inform. Theory **2** (1956), 8–19.
[62] H. S. Snevily, *The Cayley Addition Table of Z_n*, Amer. Math. Monthly 106 (1999), 184–185.
[63] Z. W. Sun, On Snevily's conjecture and restricted sumsets, to appear.
[64] M. Talagrand, *On Russo's approximate zero-one law*, Ann. Probab. 22 (1994), 1576–1587.
[65] B. L. van der Waerden, **Modern Algebra**, Julius Springer, Berlin, 1931.
[66] L. Vigneron, *Remarques sur les réseaux cubiques de classe 3 associés au probléme des quatre couleurs*, C. R. Acad. Sc. Paris, t. 223 (1946), 770–772.
[67] V. G. Vizing, *Coloring the vertices of a graph in prescribed colors* (in Russian), Diskret. Analiz. No. 29, Metody Diskret. Anal. v. Teorii Kodov i Shem 101 (1976), 3–10.

ICM 2002 · Vol. I · 137–157

Differential Complexes and Numerical Stability

Douglas N. Arnold*

Abstract

Differential complexes such as the de Rham complex have recently come to play an important role in the design and analysis of numerical methods for partial differential equations. The design of stable discretizations of systems of partial differential equations often hinges on capturing subtle aspects of the structure of the system in the discretization. In many cases the differential geometric structure captured by a differential complex has proven to be a key element, and a discrete differential complex which is appropriately related to the original complex is essential. This new geometric viewpoint has provided a unifying understanding of a variety of innovative numerical methods developed over recent decades and pointed the way to stable discretizations of problems for which none were previously known, and it appears likely to play an important role in attacking some currently intractable problems in numerical PDE.

2000 Mathematics Subject Classification: 65N12.
Keywords and Phrases: Finite element, Numerical stability, Differential complex.

1. Introduction

During the twentieth century chain complexes, their exactness properties, and commutative diagrams involving them pervaded many branches of mathematics, most notably algebraic topology and differential geometry. Recently such homological techniques have come to play an important role in a branch of mathematics often thought quite distant from these, numerical analysis. Their most significant applications have been to the design and analysis of numerical methods for the solution of partial differential equations.

Let us consider a general problem, such as a boundary value problem in partial differential equations, as an operator equation: given data f in some space Y find

*Institute for Mathematics and its Applications, University of Minnesota, 400 Church St. S.E., Minneapolis, MN 55455, USA. E-mail: arnold@ima.umn.edu

the solution u in some space X to the problem $Lu = f$. A numerical method discretizes this problem through the construction of an operator $L_h : X_h \to Y_h$ and data $f_h \in Y_h$ and defines an approximate solution $u_h \in X_h$ by the equation $L_h u_h = f_h$. Of course the numerical method is not likely to be of value unless it is consistent which means that L_h and f_h should be close to L and f in an appropriate sense.

Before we speak of solving the original problem, numerically or otherwise, we should first confront the question of whether it is well-posed. That is, given $f \in Y$, does a unique $u \in X$ exist, and, if so, do small changes of f induce small changes in Y? The analogous questions for the numerical method, whether given $f_h \in Y_h$ a unique $u_h \in X_h$ is determined by the discrete equation $L_h u_h = f_h$, and whether small changes in f_h induce small changes in u_h, is the question of *stability* of the numerical method. A common paradigm, which can be formalized in many contexts of numerical analysis, is that a method which is consistent and stable is convergent.

Well-posedness is a central issue in the theory of partial differential equations. Of course, we do not expect just any PDE problem to be well-posed. Well-posedness hinges on structure of the problem which may be elusive or delicate. Superficially small changes, for example to the sign of a coefficient or the type of boundary conditions, can certainly destroy well-posedness. The same is true for the stability of numerical methods: it often depends on subtle or elusive properties of the numerical scheme. Usually stability reflects some portion of the structure of the original problem that is captured by the numerical scheme. However in many contexts it is not enough that the numerical scheme be close to the original problem in a quantitative sense for it to inherit stability. That is, it may well happen that a consistent method for a well-posed problem is unstable. In this paper we shall see several examples where the exactness properties of discrete differential complexes and their relation to differential complexes associated with the PDE are crucial tools in establishing the stability of numerical methods. In some cases the homological arguments have served to elucidate or validate methods that had been developed over the preceding decades. In others they have pointed the way to stable discretizations of problems for which none were previously known. They will very likely play a similar role in the eventual solution of some formidable open problems in numerical PDE, especially for problems with significant geometric content, such as in numerical general relativity. As in other branches of mathematics, in numerical analysis differential complexes serve both to encode key structure concisely and to unify considerations from seemingly very different contexts.

In this paper we shall discuss only finite element methods since, of the major classes of numerical methods for PDE, they are the most amenable to rigorous analysis, and have seen the greatest use of differential complexes. But complexes have recently arisen in the study of finite differences, finite volumes, and spectral methods as well.

2. Finite element spaces

A finite element space on a domain Ω is a function space defined piecewise

by a certain assembly procedure which we now recall; cf. [7]. For simplicity, here we shall restrict to spaces of piecewise polynomials with respect to a triangulation of an n-dimensional domain by n-simplices with $n = 2$ or 3 (so implicitly we are assuming that $\Omega \subset \mathbb{R}^2$ is polygonal or $\Omega \subset \mathbb{R}^3$ is polyhedral). On each simplex T we require that there be given a function space of *shape function* W_T and a set of *degrees of freedom*, i.e., a set of linear functionals on W_T which form a basis for the dual space. Moreover, each degree of freedom is supposed to be associated with some subsimplex of some dimension, i.e., in three dimensions with a vertex, an edge, a face, or the tetrahedron itself. For a subsimplex which is shared by two simplices in the triangulation, we assume that the corresponding functionals are in one-to-one-correspondence. Then the finite element space W_h is defined as those functions on Ω whose restriction to each simplex T of the triangulation belongs to W_T and for which the corresponding degrees of freedom agree whenever a subsimplex is shared by two simplices.

The simplest example is obtained by choosing W_T to be the constant functions and taking as the only degree of freedom on T the 0th order moment $\phi \mapsto \int_T \phi(x)\, dx$ (which we associate with T itself). The resulting finite element space is simply the space of piecewise constant functions with respect to the given triangulation. Similarly we could choose $W_T = \mathbb{P}_1(T)$ (by $\mathbb{P}_p(T)$ we denote the space of polynomial functions on T of degree at most p), and take as degrees of freedom the moments of degrees 0 and also those of degree 1, $\phi \mapsto \int_T \phi(x) x_i\, dx$. Again all the degrees of freedom are associated to T itself. This time the finite element space consists of all piecewise linear functions. Of course, the construction extends to higher degrees.

A more common piecewise linear finite element space occurs if we again choose $W_T = \mathbb{P}_1(T)$, but take as degrees of freedom the maps $\phi \mapsto \phi(v)$, one associated to each vertex v. In this case the assembled finite element space consists of all *continuous* piecewise linear functions. More generally we can choose $W_T = \mathbb{P}_p(T)$ for $p \geq 1$, and associate to each vertex the evaluation degrees of freedom just mentioned, to each edge the moments on the edge of degree at most $p-2$, to each face the moments on the face of degree at most $p-3$, and to each tetrahedron the moments of degree at most $p-4$. The resulting finite element space, called the *Lagrange finite element* of degree p, consists of all continuous piecewise polynomials of degree at most p. Figure 1 shows a mesh of a two dimensional domain and a typical function in the space of Lagrange finite elements of degree 2 with respect to this mesh.

Mnemonic diagrams as in Figure 2 are often associated to finite element spaces, depicting a single element T and a marker for each degree of freedom.

Next we describe some finite element spaces that can be used to approximate vector-valued functions. For brevity we limit the descriptions to the 3-dimensional case, but supply diagrams in both 2 and 3 dimensions. Of course we may simply take the Cartesian product of three copies of one of the previous spaces. For example, the element diagrams shown on the left of Figure 3 refer to continuous piecewise linear vector fields in two and three dimensions. More interesting spaces are the *face elements* and *edge elements* essentially conceived by Raviart and Thomas [12] in two dimensions and by Nedelec [10] in three dimensions. In the lowest order

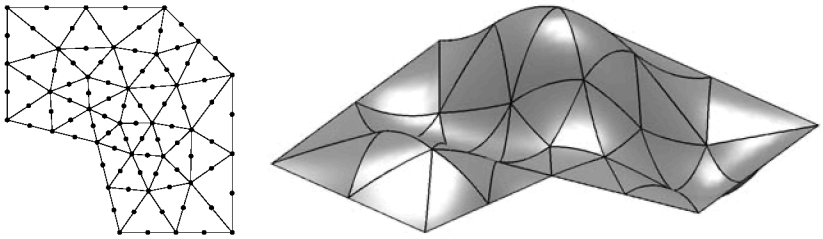


Figure 1: A mesh marked with the locations of the degrees of freedom for Lagrange finite elements of degree 2 and a typical such finite element function.

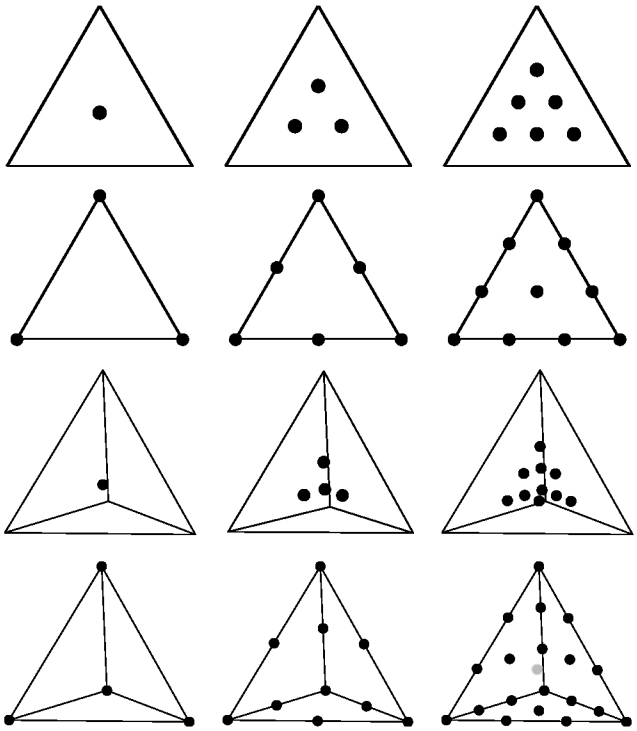


Figure 2: Element diagrams. First row: discontinuous elements of degrees 0, 1, and 2 in two dimensions. Second row: Lagrange elements of degrees 1, 2, and 3 in two dimensions. Third and fourth rows: the corresponding elements in three dimensions.

case, the face elements take as shape functions polynomial vector fields of the form $p(x) = a + bx$ where $a \in \mathbb{R}^3$, $b \in \mathbb{R}$ and $x = (x_1, x_2, x_3)$, a 4-dimensional subspace of the 12-dimensional space $\mathbb{P}_1(T, \mathbb{R}^3)$ of polynomial vector fields of degree at most

1. The degrees of freedom are taken to be the 0th order moments of the normal components on the faces of codimension 1, $p \mapsto \int_f p(x) \cdot n_f \, dx$ where f is a face and n_f the unit normal to the face. The element diagram is shown in the middle column of Figure 3. In the lowest order case the edge elements shape functions are polynomial vector fields of the form $p(x) = a + b \times x$ where $a, b \in \mathbb{R}^3$, which form a 6-dimensional subspace of $\mathbb{P}_1(T, \mathbb{R}^3)$. The degrees of freedom are the 0th order moments over the edges of the component tangent to the edge, $p \mapsto \int_e p(x) \cdot t_e \, dx$, as indicated on the right of Figure 3.

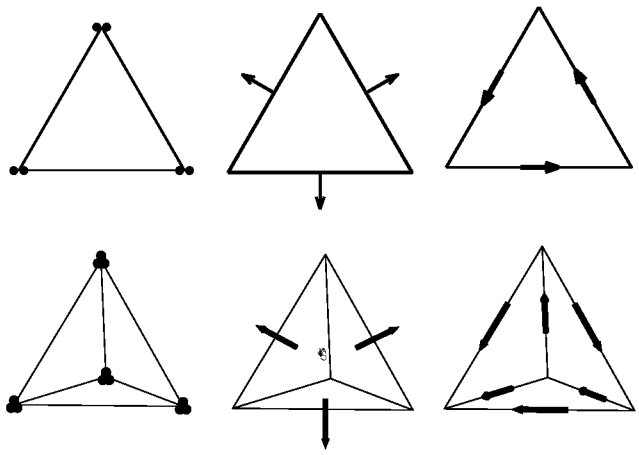


Figure 3: Element diagrams for some finite element approximations to vector fields in two and three dimensions. Multiple dots are used as markers to indicate the evaluation of all components of a vector field. Arrows are used for normal moments on codimension 1 subsimplices and for tangential components on edges. Left: continuous piecewise linear fields. Middle: face elements of lowest order. Right: edge elements of lowest order.

Each of these spaces can be generalized to arbitrarily high order. For the next higher order face space, the shape functions take the form $p(x) = a(x) + b(x)x$ where $a \in \mathbb{P}_1(T, \mathbb{R}^3)$ and $b \in \mathbb{P}_1(T)$ a linear scalar-valued polynomial. This gives a subspace of $\mathbb{P}_2(T, \mathbb{R}^3)$ of dimension 15, and the degrees of freedom are the moments of degree at most 1 of the normal components on the faces and the moments of degree 0 of all components on the tetrahedron. This element is indicated on the left of Figure 4. For the second lowest order edge space, the shape functions take the form $p(x) = a(x) + b(x) \times x$ with $a, b \in \mathbb{P}_1(T, \mathbb{R}^3)$, giving a 20-dimensional space. The degrees of freedom are the tangential moments of degree at most 1 on the edges (two per edge) and the tangential moments of degree 0 on the faces (two per face). This element is indicated on the right of Figure 4.

The choice of the shape functions and the degrees of freedom determine the smoothness of the functions belonging to the assembled finite element space. For ex-

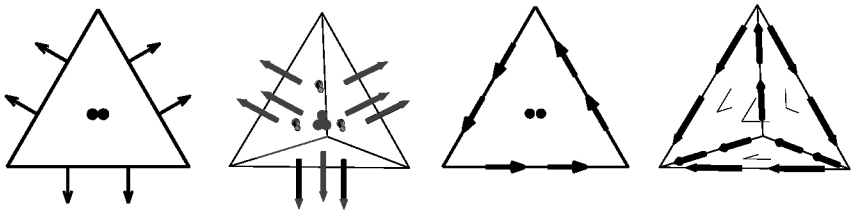


Figure 4: The face (left) and edge (right) elements of the second lowest order in 2- and 3-dimensions.

ample, the Lagrange finite element spaces of any degree belong to the Sobolev space $H^1(\Omega)$ of $L^2(\Omega)$ functions whose distributial first partial derivatives also belong to $L^2(\Omega)$ (and even to $L^infty(\Omega)$). In fact, the distributional first partial derivative of a continuous piecewise smooth function coincides with its derivative taken piecewise and so belongs to L^2. Thus the degrees of freedom we imposed in constructing the Lagrange finite elements are sufficient to insure that the assembled finite element space $W_h \subset H^1(\Omega)$. In fact more is true: for the Lagrange finite element space with shape function spaces $W_T = \mathbb{P}_p(T)$, we have

$$W_h = \{\, u \in H^1(\Omega) \,|\, u|_T \in W_T \text{ for all simplices } T \text{ of the triangulation} \,\}.$$

This says that, in a sense, the degrees of freedom impose exactly the continuity required to belong to H^1, no less and no more.

In contrast, the discontinuous piecewise polynomial spaces are subsets of $L^2(\Omega)$ but not of $H^1(\Omega)$, since their distributional first derivatives involve distributions supported on the interelement boundaries, and so do not belong to $L^2(\Omega)$.

For the vector-valued finite elements there are more possibilities. The face and edge spaces contain discontinuous functions, and so are not contained in $H^1(\Omega, \mathbb{R}^3)$. However, for vector fields belonging to one of the face spaces the normal component of the vector field does not jump across interelement boundaries, and this implies, via integration by parts, that the distributional divergence of the function coincides with the divergence taken piecewise. Thus the face spaces belong to $H(\mathrm{div}, \Omega)$, the space of L^2 vector fields on Ω whose divergence belongs to L^2. Indeed, for these spaces the degrees of freedom impose exactly the continuity of $H(\mathrm{div})$, no less or more. For the edge spaces it can be shown that the tangential components of a vector field do not jump across element boundaries, and this implies that the edge functions belong to $H(\mathrm{curl}, \Omega)$, the space of L^2 vector fields whose curl belongs to L^2. Again the degrees of freedom impose exactly the continuity needed for inclusion in $H(\mathrm{curl})$.

3. Discrete differential complexes

The de Rham complex

$$\mathbb{R} \hookrightarrow \bigwedge{}^0(\Omega) \xrightarrow{d} \bigwedge{}^1(\Omega) \xrightarrow{d} \cdots \xrightarrow{d} \bigwedge{}^n(\Omega) \to 0$$

is defined for an arbitrary smooth n-manifold Ω. Here $\bigwedge^k(\Omega)$ denotes the space of differential k-forms on Ω, i.e., for $\omega \in \bigwedge^k(\Omega)$ and $x \in \Omega$, $\omega(x)$ is an alternating k-linear map on the tangent space $T_x\Omega$. The operators $d : \bigwedge^k(\Omega) \to \bigwedge^{k+1}(\Omega)$ denote exterior differentiation. This is is a complex in that the composition of two exterior differentiations always vanishes. Moreover, and if the manifold is topologically trivial, then it is exact.

If Ω is a domain in $\mathbb{R}^3$, then we may identify its tangent space at any point with $\mathbb{R}^3$. Using the Euclidean inner product, the space of linear maps on $\mathbb{R}^3$ may be identified by $\mathbb{R}^3$ as usual, so $\bigwedge^1(\Omega)$ may be identified with the space $C^\infty(\Omega, \mathbb{R}^3)$ of smooth vector fields on Ω. Moreover, the space of alternating bilinear maps on $\mathbb{R}^3$ may be identified with $\mathbb{R}^3$ by associating to a vector u the alternating bilinear map $(v, w) \mapsto \det(u|v|w)$. Thus we have an identification of $\bigwedge^2(\Omega)$ with $\mathbb{R}^3$ as well. Finally the only alternating trilinear maps on $\mathbb{R}^3$ are given by multiples of the determinant map $(u, v, w) \mapsto c\det(u|v|w)$, and so we may identify $\bigwedge^3(\Omega)$ with $C^\infty(\Omega)$. In terms of such *proxy fields*, the de Rham complex becomes

$$\mathbb{R} \hookrightarrow C^\infty(\Omega) \xrightarrow{\text{grad}} C^\infty(\Omega, \mathbb{R}^3) \xrightarrow{\text{curl}} C^\infty(\Omega, \mathbb{R}^3) \xrightarrow{\text{div}} C^\infty(\Omega, \mathbb{R}) \to 0. \tag{3.1}$$

Alternatively we may consider L^2-based forms and the sequence becomes

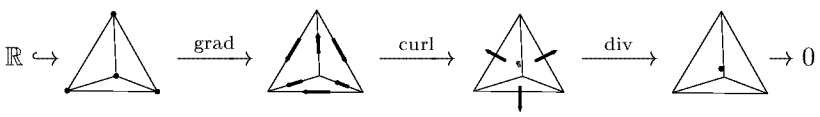
$$\mathbb{R} \hookrightarrow H^1(\Omega) \xrightarrow{\text{grad}} H(\text{curl}, \Omega) \xrightarrow{\text{curl}} H(\text{div}, \Omega) \xrightarrow{\text{div}} L^2(\Omega, \mathbb{R}) \to 0.$$

The finite element spaces constructed above allow us to form discrete analogues of the de Rham complex. Given some triangulation of $\Omega \subset \mathbb{R}^3$, let W_h denote the space of continuous piecewise linear finite elements, Q_h the lowest order edge element space, S_h the lowest order face element space, and V_h the space of piecewise constants. Then $\text{grad}\, W_h \subset Q_h$ (since Q_h contains all piecewise constant vector fields belonging to $H(\text{curl})$ and the gradient of a continuous piecewise linear is certainly such a function), $\text{curl}\, Q_h \subset S_h$ (since S_h contains all piecewise constant vector fields belonging to $H(\text{curl})$), and $\text{div}\, S_h \subset V_h$. Thus we have the discrete differential complex

$$\mathbb{R} \hookrightarrow W_h \xrightarrow{\text{grad}} Q_h \xrightarrow{\text{curl}} S_h \xrightarrow{\text{div}} V_h \to 0. \tag{3.2}$$

This differential complex captures the topology of the domain to the same extent as the de Rham complex. In particular, if the domain is topologically trivial, then the sequence is exact.

It is convenient to abbreviate the above statement using the element diagrams introduced earlier. Thus we will say that the following complex is exact:



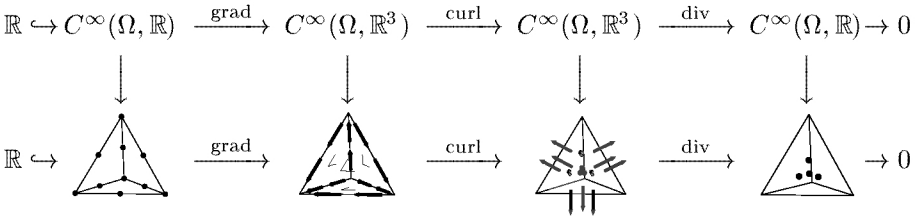
By this we mean that if we assemble finite element spaces W_h, Q_h, S_h, and V_h using the indicated finite elements and a triangulation of a topologically trivial domain, then the corresponding discrete differential complex (3.2) is exact.

There is another important relationship between the de Rham complex (3.1) and the discrete complex (3.2). The defining degrees of freedom determine projections $\Pi_h^W : C^\infty(\Omega) \to W_h$, $\Pi_h^Q : C^\infty(\Omega, \mathbb{R}^3) \to Q_h$, and so on. In fact Π_h^W is just the usual interpolant, Π_h^V is the L^2-projection into the piecewise constants, and the projections Π_h^Q and Π_h^S onto the edge and face elements are determined by the maintenance of the appropriate moments. It can be checked, based on Stokes theorem, that the following diagram commutes.

$$\begin{array}{ccccccccc}
\mathbb{R} \hookrightarrow & C^\infty(\Omega,\mathbb{R}) & \xrightarrow{\text{grad}} & C^\infty(\Omega,\mathbb{R}^3) & \xrightarrow{\text{curl}} & C^\infty(\Omega,\mathbb{R}^3) & \xrightarrow{\text{div}} & C^\infty(\Omega,\mathbb{R}) & \to 0 \\
& \downarrow \Pi_h^W & & \downarrow \Pi_h^Q & & \downarrow \Pi_h^S & & \downarrow \Pi_h^V & \\
\mathbb{R} \hookrightarrow & W_h & \xrightarrow{\text{grad}} & Q_h & \xrightarrow{\text{curl}} & S_h & \xrightarrow{\text{div}} & V_h & \to 0
\end{array} \tag{3.3}$$

The finite element spaces appearing in this diagram, with one degree of freedom for each vertex for W_h, for each edge for Q_h, for each face for S_h, and for each simplex for V_h, are highly geometrical. In fact, recalling the identifications between fields and differential forms, we may view these spaces as spaces of piecewise smooth differential forms. They were in fact first constructed in this context, without any thought of finite elements or numerical methods, by Whitney [13]. The spaces were reinvented, one-by-one, as finite element spaces in response to the needs of various numerical problems, and the properties which are summarized in the commutative diagram above were slowly rediscovered as needed to analyze the resulting numerical methods. The connection between low order edge and face finite elements and Whitney forms was first realized by Bossavit [5].

Analogous statements hold for higher order Lagrange, edge, face, and discontinuous finite elements. For example, the following diagram commutes and has exact rows:



We shall see many other discrete differential complexes below.

4. Stability of Galerkin methods

Consider first the solution of the Dirichlet problem for Poisson's equation on a domain in $\mathbb{R}^n$:

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega.$$

The solution can be characterized as the minimizer of the energy functional

$$\mathcal{E}(u) := \frac{1}{2}\int_\Omega |\operatorname{grad} u(x)|^2\, dx - \int_\Omega f(x)u(x)\, dx$$

over the Sobolev space $\mathring{H}^1(\Omega)$ (consisting of $H^1(\Omega)$ functions vanishing on $\partial\Omega$), or as the solution of the weak problem: find $u \in \mathring{H}^1(\Omega)$ such that

$$\int_\Omega \operatorname{grad} u(x) \cdot \operatorname{grad} v(x)\, dx = \int_\Omega f(x)v(x)\, dx \quad \text{for all } v \in \mathring{H}^1(\Omega).$$

We may define an approximate solution u_h by minimizing the Dirichlet integral over a finite dimensional subspace W_h of $\mathring{H}^1(\Omega)$; this is the classical Ritz method. Equivalently, we may use the Galerkin method, in which $u_h \in W_h$ is determined by the equations

$$\int_\Omega \operatorname{grad} u_h(x) \cdot \operatorname{grad} v(x)\, dx = \int_\Omega f(x)v(x)\, dx \quad \text{for all } v \in W_h.$$

After choice of a basis in W_h this leads to a system of linear algebraic equations, and u_h is computable.

Let T_h denote the discrete solution operator $f \mapsto u_h$. Then it is easy to check that T_h is bounded as a linear operator from $H^{-1}(\Omega) := \mathring{H}^1(\Omega)^*$ to $\mathring{H}^1(\Omega)$ by a constant that depends only on the domain Ω (and, in particular, doesn't increase if the space W_h is enriched). This says that the Galerkin method is *stable*. A consequence is the *quasioptimality estimate*

$$\|u - u_h\|_{H^1} \le c \inf_{v \in W_h} \|u - v\|_{H^1}, \tag{4.4}$$

for some constant c depending only on the domain Ω. Note that there is no restriction on the subspace W_h to obtain this estimate. Galerkin's method for a coercive elliptic problem is always stable and convergence depends only on the approximation properties of the subspace. A natural choice for W_h is the Lagrange finite element space of some degree p with respect to some regular simplicial mesh of maximal element size h, in which case Galerkin's method is a standard finite element method. In this case the right hand side of (4.4) is $O(h^p)$ provided that u is sufficiently smooth.

Next consider the related eigenvalue problem, which arises in the determination of the fundamental frequencies of a drum. That is, we seek standing wave solutions $w(x,t)$ to the wave equation on some bounded domain $\Omega \subset \mathbb{R}^2$ which vanish on $\partial\Omega$. Assuming that the tension and density of the drum membrane are unity, these solutions have the form $w(x,t) = \alpha \cos(\sqrt{\lambda}t)u(x) + \beta \sin(\sqrt{\lambda}t)u(x)$ where α and β are constants and u and λ satisfy the eigenvalue problem

$$-\Delta u = \lambda u \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega.$$

The eigenvalues λ form a sequence of positive numbers tending to infinity. The numbers $\sqrt{\lambda}/(2\pi)$ are the fundamental frequencies of the drum and the functions u give the corresponding fundamental modes.

The eigenvalues and eigenfunctions are characterized variationally as the critical values and critical points of the Rayleigh quotient

$$\mathcal{R}(u) = \frac{\int_\Omega |\operatorname{grad} u(x)|^2\, dx}{\int_\Omega |u(x)|^2\, dx},$$

defined for nonzero u belonging to the Sobolev space $\mathring{H}^1(\Omega)$. The classical Rayleigh-Ritz method for the approximation of eigenvalue problems determines approximate eigenvalues λ_h and eigenfunctions u_h as the critical values and points of the restriction of $\mathcal{R}$ to the nonzero elements of some finite dimensional subspace W_h of $\mathring{H}^1(\Omega)$. Equivalently, we can write the eigenvalue problem in weak form: find $\lambda \in \mathbb{R}$ and nonzero $u \in \mathring{H}^1(\Omega)$ such that

$$\int_\Omega \operatorname{grad} u(x) \cdot \operatorname{grad} v(x)\, dx = \lambda \int_\Omega u(x)v(x)\, dx \quad \text{for all } v \in \mathring{H}^1(\Omega). \tag{4.5}$$

The Galerkin approximation of the eigenvalue problem, which is equivalent to the Rayleigh-Ritz method, seeks $\lambda_h \in \mathbb{R}$ and nonzero $u_h \in W_h$ such that

$$\int_\Omega \operatorname{grad} u_h(x) \cdot \operatorname{grad} v(x)\, dx = \lambda_h \int_\Omega u_h(x)v(x)\, dx \quad \text{for all } v \in W_h. \tag{4.6}$$

We now discuss the convergence of this method. Let λ denote the jth eigenvalue of the problem (4.5). In the interest of simplicity we assume that λ is a simple eigenvalue, so the corresponding eigenfunction u is uniquely determined up to sign by the normalization $\|u\|_{H^1} = 1$. Similarly let λ_h and u_h denote the jth eigenvalue of (4.6). It can then be proved (see, e.g., [3] for much more general results) that there exists a constant c such that

$$\|u - u_h\|_{H^1} \le c \inf_{v \in W_h} \|u - v\|_{H^1}, \quad |\lambda - \lambda_h| \le c\|u - u_h\|^2_{H^1}. \tag{4.7}$$

In short, the eigenfunction approximation is quasioptimal and the eigenvalue error is bounded by the square. Again there is no restriction on the space W_h.

Figure 5 reports on the computation of the eigenvalues of the Laplacian on an elliptical domain of aspect ratio 3 using Lagrange finite elements of degree 1.

Now consider an analogous problem, the computation of the resonant frequencies of an electromagnetic cavity occupying a region $\Omega \subset \mathbb{R}^3$. In this case we wish to find standing wave solutions of Maxwell's equations. If we take the electric permittivity and the magnetic permeability to be unity and assume a lossless cavity with perfectly conducting boundary, we are led to the following eigenvalue problem for the electric field: find nonzero $E : \Omega \to \mathbb{R}^3$, $\lambda \in \mathbb{R}$ such that

$$\operatorname{curl}\operatorname{curl} E = \lambda E, \quad \operatorname{div} E = 0 \text{ in } \Omega, \quad E \times n = 0 \text{ on } \partial\Omega. \tag{4.8}$$

This is again an elliptic eigenvalue problem and the eigenvalues form a sequence of positive numbers tending to infinity. The divergence constraint is nearly redundant in this eigenvalue problem. Indeed if $\operatorname{curl}\operatorname{curl} E = \lambda E$ for $\lambda > 0$, then $\operatorname{div} E = \lambda^{-1} \operatorname{div}\operatorname{curl}\operatorname{curl} E = 0$ since the divergence of a curl vanishes. Thus the eigenvalue problem

$$\operatorname{curl}\operatorname{curl} E = \lambda E \text{ in } \Omega, \quad E \times n = 0 \text{ on } \partial\Omega, \tag{4.9}$$

has the same eigenvalues and eigenfunctions as (4.8) except that it also admits $\lambda = 0$ as an eigenvalue, and the corresponding eigenspace is infinite-dimensional (it contains the gradients of all smooth functions vanishing on the boundary of Ω). The

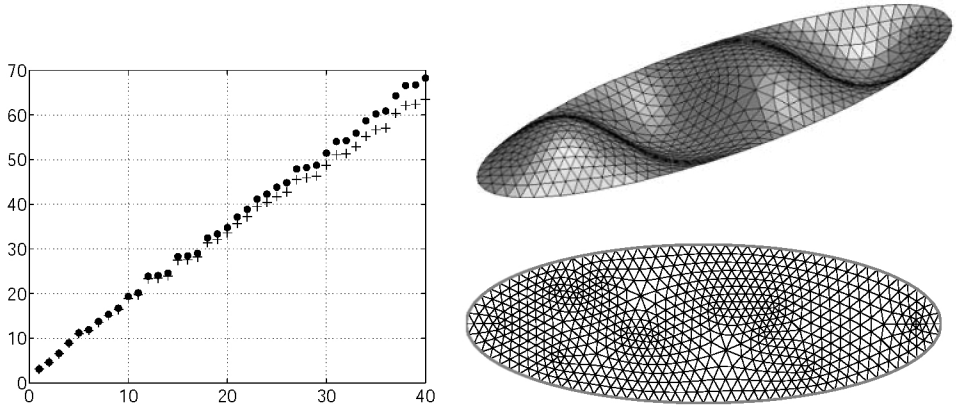


Figure 5: The point plot shows the first 40 eigenvalues computed with piecewise linear finite elements with respect to the triangulation shown (•) versus the exact eigenvalues (+). The surface plot shows the computed eigenfunction associated to the fourth eigenvalue. The mesh has 737 vertices, of which 641 are interior, and 1,376 triangles.

eigenvalues and eigenfunctions are now critical points and values of the Rayleigh quotient

$$\mathcal{R}(E) = \frac{\int_\Omega |\operatorname{curl} E(x)|^2 dx}{\int_\Omega |E(x)|^2 dx},$$

over the space of nonzero fields E in $\mathring{H}(\operatorname{curl}, \Omega)$, which is defined to be the space of functions for which both the above integrals exist and are finite and which have vanishing tangential component on the boundary (i.e., $E \times n = 0$ on $\partial\Omega$).

In Figure 6 we show the result of approximating a two-dimensional version of this eigenvalue problem using the Rayleigh-Ritz method or, equivalently, the Galerkin method with continuous piecewise linear vector fields on Ω whose tangential components vanish on the boundary (the first element depicted in Figure 3). For Ω we take a square of side length π, in which case the nonzero eigenvalues are known to be all numbers of the form $\lambda = m^2 + n^2$ with $0 \leq m, n \in \mathbb{Z}$ not both zero, and the corresponding eigenfunctions are $E = (\sin my, \sin nx)$. For the mesh pictured, the finite element space has dimension 290. We find that 73 of the 290 computed eigenvalues are between 0 and 10 and that they have no tendency to cluster near the integers $1, 1, 2, 4, 4, 5, 5, 8, 9, 9$ which are the exact eigenvalues between 0 and 10. Thus this numerical method is useless: the computed eigenvalues bear no relation to the true eigenvalues! The analogue of (4.7) is surely not true.

If instead we choose the lowest order edge elements as the finite element space (Figure 3, top right), we get very different results. Using the same mesh, the edge

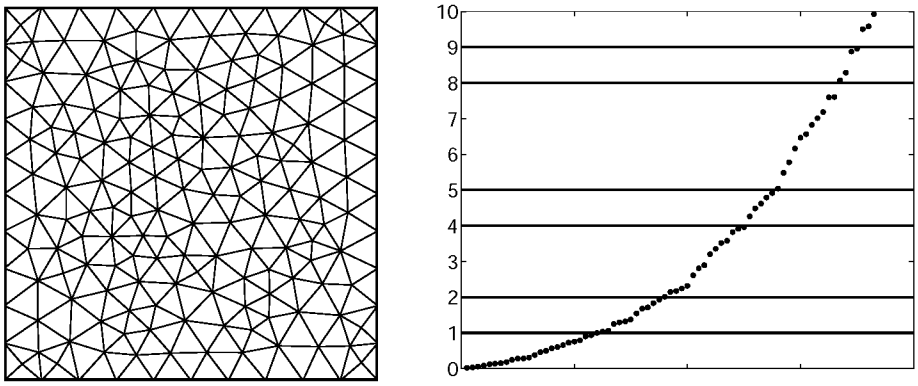


Figure 6: The plot shows the first 73 eigenvalues computed with piecewise linear finite elements for the resonant cavity problem on the square using the mesh shown. They bear no relation to the exact eigenvalues, 1, 1, 2, 4, 4, ..., indicated by the horizontal lines.

finite element space has dimension 472. It turns out that 145 of the computed eigenvalues are zero (to within round-off), and the subsequent eigenvalues are 0.9998, 0.9999, 2.0023, 3.9968, 4.0013, ..., i.e., excellent approximations of the exact eigenvalues. See Figure 7.

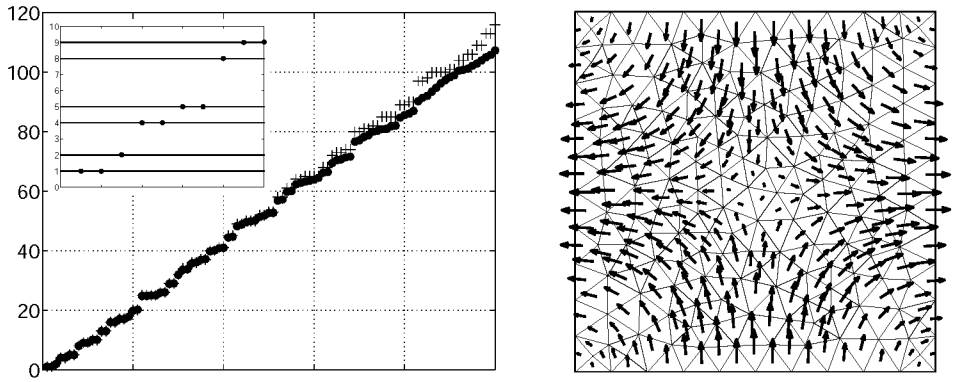


Figure 7: The first plot shows the first 100 positive eigenvalues for the resonant cavity problem on the square computed with lowest order edge elements using the mesh of Figure 6. The error in the first 54 eigenvalues is below 2%. The inset focuses on the first 10 eigenvalues, for which the error is less than 0.25%. The second plot shows the vector field associated to the third positive eigenvalue.

The striking difference between the behavior of the continuous piecewise linear finite elements and the edge elements for the resonant cavity problem is a question of stability. We shall return to this below, after examining stability in a simpler context.

5. Stability of mixed formulations

Consider now the Dirichlet problem

$$-\operatorname{div} C \operatorname{grad} u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega,$$

where Ω is a domain in $\mathbb{R}^3$ and the coefficient C is a symmetric positive definite matrix at each point. We may again characterize u as a minimizer of the energy functional

$$u \mapsto \frac{1}{2}\int C \operatorname{grad} u \cdot \operatorname{grad} u dx - \int f u \, dx$$

and use the Ritz method. This procedure is always stable.

However, for some purposes it is preferable to work with the equivalent first order system

$$\sigma = C \operatorname{grad} u, \quad -\operatorname{div} \sigma = f. \tag{5.10}$$

The pair (σ, u) is then characterized variationally as the unique critical point of the functional

$$\mathcal{L}(\sigma, u) = \int_\Omega (\frac{1}{2} C^{-1}\sigma \cdot \sigma + u \operatorname{div} \sigma) dx - \int_\Omega f u \, dx \tag{5.11}$$

over $H(\operatorname{div}, \Omega) \times L^2(\Omega)$. Note that (σ, u) is a saddle-point of $\mathcal{L}$, not an extremum. Numerical discretizations based on such saddle-point variational principles are called *mixed methods.*

It is worth interpreting the system (5.10) in the language of differential forms, because this brings some insight. The function u is a 0-form, and the operation $u \mapsto \operatorname{grad} u$ is just exterior differentiation. The vector field σ is a proxy for a 2-form and the operation $\sigma \mapsto \operatorname{div} \sigma$ is again exterior differentiation. The loading function f is the proxy for a 3-form. Since $\operatorname{grad} u$ is the proxy for a 1-form, it must be that the operation on differential forms that corresponds to multiplication by C takes 1-forms to 2-forms. In fact, if we untangle the identifications, we find that multiplication by C is a Hodge star operation. A Hodge star operator defines an isomorphism of $\bigwedge^k(\Omega)$ onto $\bigwedge^{3-k}(\Omega)$. To determine a particular such operator, we must define an inner product on the tangent space $\mathbb{R}^3$ at each point of Ω. The positive definite matrix C does exactly that. Many of the partial differential equations of mathematical physics admit similar interpretations in terms of differential forms. For a discussion of this in the context of discretization, see [9].

A natural approach to discretization of the mixed variational principle is to choose subspaces $S_h \subset H(\operatorname{div}, \Omega)$, $V_h \subset L^2(\Omega)$ and seek a critical point $(\sigma_h, u_h) \in S_h \times V_h$. This is of course equivalent to a Galerkin method and leads to a system of linear algebraic equations. However in this case, *stability is not automatic.* It can

happen that the discrete system is singular, or more commonly, that the norm of the discrete solution operator grows unboundedly as the mesh is refined.

In a fundamental paper, Brezzi [6] established two conditions that together are sufficient (and essentially necessary) for stability. Brezzi's theorem applied to a wide class of saddle-point problems, but for simplicity we will state the stability conditions for the saddle-point problem associated to the functional (5.11).

(S1) There exists $\gamma_1 > 0$ such that

$$\int_\Omega C^{-1}\tau \cdot \tau \, dx \geq \gamma_1 \|\tau\|^2_{H(\mathrm{div})},$$

for all $\tau \in S_h$ such that $\int \operatorname{div} \tau \, v \, dx = 0$ for all $v \in V_h$.

(S2) There exists $\gamma_2 > 0$ such that for all $v \in V_h$ there exists nonzero $\tau \in S_h$ satisfying

$$\int_\Omega v \operatorname{div} \tau \, dx \geq \gamma_2 \|v\|_{L^2} \|\tau\|_{H(\mathrm{div})}.$$

Theorem (Brezzi) *If the stability conditions* (S1) *and* (S2) *are satisfied, then* $\mathcal{L}$ *admits a unique critical point* (σ_h, u_h) *over* $S_h \times V_h$*, the solution operator* $f \mapsto (\sigma_h, u_h)$ *is bounded* $L^2(\Omega) \to H(\mathrm{div}, \Omega) \times L^2(\Omega)$*, and the quasioptimal estimate*

$$\|\sigma - \sigma_h\|_{H(\mathrm{div})} + \|u - u_h\|_{L^2} \leq c \inf_{(\tau,v)\in S_h\times V_h} (\|\sigma - \tau\|_{H(\mathrm{div})} + \|u - v\|_{L^2})$$

holds with c *depending on* γ_1 *and* γ_2.

The stability conditions of Brezzi strongly limit the choice of the mixed finite element spaces S_h and V_h. Condition (S1) is satisfied if the indicated functions $\tau \in S_h$, those whose divergence is orthogonal to V_h, are in fact divergence-free. (In practice, this is nearly the only way it is satisfied.) This certainly holds if $\operatorname{div} S_h \subset V_h$, and so such as inclusion is a common design principle of mixed finite element spaces. On the other hand, condition (S2) is most easily satisfied if $\operatorname{div} S_h \supset V_h$, because in this case, given $v \in V_h$, we can choose $\tau \in S_h$ with $\operatorname{div} \tau = v$, so $\int_\Omega v \operatorname{div} \tau \, dx = \|v\|^2_{L^2}$, and the second condition will be satisfied as long as we can insure that $\|\tau\|_{H(\mathrm{div})} \leq \gamma_2^{-1} \|v\|_{L^2}$. In short, we need to know that div maps S_h onto V_h and that $\operatorname{div}|_{S_h}$ admits a bounded one-sided inverse.

The face elements of Raviart-Thomas and Nedelec were designed to satisfy both these conditions. Specifically, let S_h again denote the space of face elements of lowest degree (whose element diagram is shown in the middle of the second row of Figure 3), and V_h the space of piecewise constants.[1] We know that $S_h \subset H(\mathrm{div}, \Omega)$ so these elements are admissable for the mixed variational principle. Moreover, we have $\operatorname{div} S_h \subset V_h$, so (S1) holds.

To verify (S2), we refer to the commutative diagram (3.3). Given $v \in V_h$, we can solve the Poisson equation $\Delta\phi = v$ and take $\sigma = \operatorname{grad} \phi$ to obtain a function

[1] It may seem odd to seek u_h in V_h, a space of discrete 3-forms, rather than in a space of 0-forms, since u is a 0-form. The resolution is through a Hodge star operator, this time formed with respect to the Euclidean inner product on $\mathbb{R}^3$. In the mixed method u_h is a discrete 3-form, approximating the image of u under this star operator.

with $\operatorname{div}\sigma = v$ and $\|\sigma\|_{H^1} \leq C\|v\|_{L^2}$. Now let $\tau = \Pi_h^S \sigma \in S_h$. Then

$$\operatorname{div}\tau = \operatorname{div}\Pi_h^S\sigma = \Pi_h^V \operatorname{div}\sigma = \Pi_h^V v = v,$$

where we have used the commutativity and the fact that $v \in V_h$. Moreover $\|\tau\|_{H(\operatorname{div})} \leq c\|\sigma\|_{H^1} \leq c'\|v\|_{L^2}$, where we used the boundedness of Π_h^S on $H^1(\Omega, \mathbb{R}^3)$. This shows that $\operatorname{div} V_h = S_h$ and establishes a bound on the one-sided inverse, and so verifies (S2). Of course, the same argument shows the stability of a mixed method based on higher order face elements as well.

Thus we see that the stability of the mixed finite element method depends on the properties of the spaces V_h and S_h encoded in the rightmost square of the commutative diagram (3.3).

Now let us return to the resonant cavity eigenvalue problem (4.9) for which we explored the Galerkin method: find $\lambda_h \in \mathbb{R}$, $0 \neq E_h \in Q_h$ such that

$$\int_\Omega \operatorname{curl} E_h \cdot \operatorname{curl} F \, dx = \lambda_h \int_\Omega E_h \cdot F \, dx \quad \text{for all } F \in Q_h. \tag{5.12}$$

We saw that if $Q_h \subset \mathring{H}(\operatorname{curl}, \Omega)$ is taken to be a space of edge elements this method gives good results in that the positive eigenvalues of the discrete problem are good approximations for the positive eigenvalues of the continuous problem. However, the simple choice of Lagrange finite elements did not give good results. We now explain the good performance of the edge elements based on the middle square of the commutative diagram (3.3). Following Boffi et. al [4] we set $P_h = \operatorname{curl} Q_h$ and introduce the following mixed discrete eigenvalue problem: find $\lambda_h \in \mathbb{R}$, $0 \neq (E_h, p_h) \in Q_h \times P_h$ such that

$$\int_\Omega E_h \cdot F \, dx + \int_\Omega \operatorname{curl} F \cdot p_h \, dx = 0 \quad \text{for all } F \in Q_h, \tag{5.13}$$

$$\int_\Omega \operatorname{curl} E_h \cdot q \, dx = -\lambda_h \int_\Omega p_h \cdot q \, dx \quad \text{for all } q \in P_h. \tag{5.14}$$

It is then easy to verify that if λ_h, E_h is a solution to (5.12) with $\lambda_h > 0$, then λ_h, $(E_h, \lambda_h^{-1} \operatorname{curl} E_h)$ is a solution to (5.13), and if λ_h, (E_h, p_h) is a solution to (5.13) then $\lambda_h > 0$ and λ_h, E_h is a solution to (5.12). In short, the two problems are equivalent except that the former admits a zero eigenspace which the mixed formulation suppresses. As explained in [4], the accuracy of the mixed eigenvalue problem (5.13) hinges on the stability of the corresponding mixed source problem. This is a saddle-point problem of the sort studied by Brezzi, and so stability depends on conditions analogous to (S1) and (S2). The proof of these conditions in case Q_h is the space of edge elements follows, as in the preceding stability verification, from surjectivity and commutativity properties encoded in the diagram (3.3).

The diagram can also be used to explain the zero eigenspace computed with edge elements. Recall that in the case of the mesh shown in Figure 6, this space had dimension 145. In fact, this eigenspace is simply the null space of the curl operator restricted to Q_h. Referring again to the commutative diagram (3.3), this is the gradient of the space W_h of linear Lagrange elements vanishing on the boundary. Its dimension is therefore exactly the number of interior nodes of the mesh.

6. The elasticity complex

Let $\mathbb{S}$ denote the space of 3×3 symmetric matrices. Given a volumetric loading density $f : \Omega \to \mathbb{R}^3$, the system of linearized elasticity determines the displacement field $u : \Omega \to \mathbb{R}^3$ and the stress field $\sigma : \Omega \to \mathbb{S}$ induced in the elastic domain Ω by the equations

$$\sigma = C\,\epsilon\, u, \quad -\operatorname{div}\sigma = f,$$

together with boundary conditions such as $u = 0$ on $\partial\Omega$. Here $\epsilon\, u$ is the symmetric part of the matrix $\operatorname{grad} u$, and the elasticity tensor $C : \mathbb{S} \to \mathbb{S}$ is a symmetric positive definite linear operator describing the particular elastic material, possibly varying from point to point.

The solution (σ, u) may be characterized variationally as a saddle-point of the Hellinger-Reissner functional

$$\mathcal{L}(\sigma, u) = \int_\Omega (\frac{1}{2} C^{-1}\sigma : \sigma + u \cdot \operatorname{div}\sigma) dx - \int_\Omega f \cdot u\, dx \tag{6.15}$$

over $H(\operatorname{div}, \Omega, \mathbb{S}) \times L^2(\Omega, \mathbb{R}^2)$ (i.e., σ is sought in the space of square-integrable symmetric-matrix-valued functions whose divergence by rows is square-integrable, and u is sought among all square-integrable vector fields).

For a mixed finite element method, we need to specify finite element subspaces $S_h \subset H(\operatorname{div}, \Omega, \mathbb{S})$ and $V_h \subset L^2(\Omega, \mathbb{R}^2)$ and restrict the domain of the variational problem. Of course the spaces must be carefully designed if the mixed method is to be stable: the analogues of the stability conditions (S1) and (S2) must be satisfied. The functional (6.15) is quite similar in appearance to (5.11) and so it might be expected that the mixed finite elements developed for the latter (the face elements for σ and discontinuous elements for u) could be adapted to the case of elasticity. In fact, the requirement of symmetry of the stress tensor and, correspondingly, the replacement of the gradient by the symmetric gradient, changes the structure significantly. Four decades of searching for mixed finite elements for elasticity beginning in the 1960s did not yield any stable elements with polynomial shape functions.

Using discrete differential complexes, R. Winther and the author recently developed the first such elements for elasticity problems in two dimensions [1]. (The three-dimensional case remains open.) For elasticity, the displacement and stress fields cannot be naturally interpreted as differential forms and the relevant differential complex is not the de Rham complex. In three dimensions it is instead the *elasticity complex*:

$$\mathbb{T} \hookrightarrow C^\infty(\Omega, \mathbb{R}^3) \xrightarrow{\epsilon} C^\infty(\Omega, \mathbb{S}) \xrightarrow{J} C^\infty(\Omega, \mathbb{S}) \xrightarrow{\operatorname{div}} C^\infty(\Omega, \mathbb{R}^3) \to 0.$$

Here the operator J is a *second order differential operator* which acts on a symmetric matrix field by first replacing each row with its curl and then replacing each column with its curl to obtain another symmetric matrix field. The resolved space $\mathbb{T}$ is the six-dimensional space of infinitesimal rigid motions, i.e., the same space of linear polynomials $a + b \times x$ which arose as the shape functions for the lowest order edge elements. If the domain Ω is topologically trivial, this complex is exact. Although

it involves a second order differential operator, and so looks quite different from the de Rham complex, Eastwood [8] recently pointed out that it can be derived from the de Rham complex via a general construction known as the Bernstein-Gelfand-Gelfand resolution.

In two dimensions the elasticity complex takes the form

$$\mathbb{P}_1 \hookrightarrow C^\infty(\Omega) \xrightarrow{J} C^\infty(\Omega, \mathbb{S}) \xrightarrow{\text{div}} C^\infty(\Omega, \mathbb{R}^2) \to 0,$$

where now the second order differential operator is

$$J = \begin{pmatrix} \dfrac{\partial^2}{\partial x_2^2} & -\dfrac{\partial^2}{\partial x_1 \partial x_2} \\ -\dfrac{\partial^2}{\partial x_1 \partial x_2} & \dfrac{\partial^2}{\partial x_1^2} \end{pmatrix}.$$

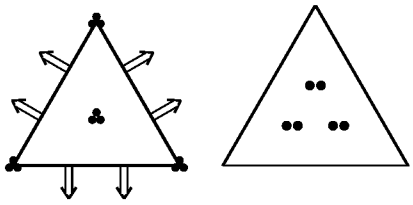


Figure 8: Element diagram for the new mixed finite elements for elasticity, lowest order case.

In the lowest order case, the finite elements we introduced in [1], for which the element diagrams can be seen in Figure 8, use discontinuous piecewise linear vector fields for the displacement field and a piecewise polynomial space which we shall now describe for the stress field. The shape functions on an arbitrary triangle T are given by

$$S_T = \{ \tau \in \mathbb{P}_3(T, \mathbb{S}) \mid \text{div}\, \tau \in \mathbb{P}_1(T, \mathbb{R}^2) \},$$

which is a 24-dimensional space consisting of all quadratic symmetric matrix fields on T together with the divergence-free cubic fields. The degrees of freedom are

- the values of three components of $\tau(x)$ at each vertex x of T (9 degrees of freedom)
- the values of the moments of degree 0 and 1 of the two components of τn on each edge e of T (12 degrees of freedom)
- the value of the three components of the moment of degree 0 of τ on T (3 degrees of freedom)

Note that these degrees of freedom are enough to ensure continuity of τn across element faces, and so will furnish a finite element subspace of $H(\text{div}, \Omega, \mathbb{S})$. The continuity is not however, the minimal needed for inclusion in $H(\text{div})$. The degrees of freedom also enforce continuity at the vertices, which is not required for

membership in $H(\mathrm{div})$. For various reasons, it would be useful to have a mixed finite element for elasticity that does not use vertex degrees of freedom. But, as we remark below, this is not possible if we restrict to polynomial shape functions.

In order to have a well-defined finite element, we must verify that the 24 degrees of freedom form a basis for the dual space of S_T. We include this verification since it illustrates an aspect of the role of the elasticity complex. Since $\dim S_T = 24$, we need only show that if all the degrees of freedom vanish for some $\tau \in S_T$, then $\tau = 0$. Now τn varies cubically along each edge, vanishes at the endpoints, and has vanishing moments of degree 0 and 1. Therefore $\tau n \equiv 0$. Letting $v = \operatorname{div} \tau$, a linear vector field on T, we get by integration by parts that

$$\int_T v^2 \, dx = -\int_T \tau : \epsilon v \, dx + \int_{\partial T} \tau n \cdot v \, ds = 0$$

since the integral of τ vanishes as well as τn. Thus τ is divergence-free. In view of the exactness of the elasticity complex, $\tau = Jq$ for some smooth function q. Since all the second partial derivatives of q belong to $\mathbb{P}_3(T)$, $q \in \mathbb{P}_5(T)$. Adjusting by an element of $\mathbb{P}_1(T)$ (the null space of J), we may take q to vanish at the vertices. Now $\partial^2 q/\partial s^2 = \tau n \cdot n = 0$ on each edge, whence q is identically zero on ∂T. This implies that the gradient of q vanishes at the vertices. Since $\partial^2 q/\partial s \partial n = -\tau n \cdot t = 0$ on each edge (with t a unit vector tangent to the edge), we conclude that $\partial q/\partial n$ vanishes identically on ∂T as well. Since q has degree at most 5, it must vanish identically.

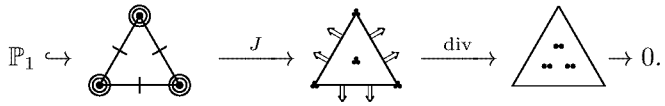
Let $\Pi_h^S : C^\infty(\Omega, \mathbb{S}) \to S_h$ denote the projection associated with the supplied degrees of freedom, and $\Pi_h^V : C^\infty(\Omega, \mathbb{R}^2) \to V_h$ the L^2-projection. For any triangle T, $\tau \in C^\infty(\Omega, \mathbb{S})$, and $v \in \mathbb{P}_1(\Omega, \mathbb{R}^2)$, we have

$$\int_T \operatorname{div}(\tau - \Pi_h^S \tau) \cdot v \, dx = -\int_T (\tau - \Pi_h^S \tau) : \epsilon v \, dx + \int_{\partial T} (\tau - \Pi_h^S \tau) n \cdot v \, ds.$$

The degrees of freedom entering the definition of Π_h^S ensure that the right hand side vanishes, and from this we obtain the commutativity $\operatorname{div} \Pi_h^S \tau = \Pi_h^V \operatorname{div} \tau$ which is essential for stability. (Actually a technical difficulty arises here, since Π_h^S as given is not bounded on $H^1(\Omega, \mathbb{S})$. See [1] for the resolution.) Note that, by their definitions, $\operatorname{div} S_h \subset V_h$ and, using the commutativity, we have $\operatorname{div} S_h = V_h$, i.e., $S_h \xrightarrow{\operatorname{div}} V_h \to 0$ is exact. To complete this to a discrete analogue of the elasticity complex, we define Y_h to be the inverse image of S_h under J. Then Y_h is exactly the space of C^1 piecewise quintic polynomials which are C^2 at the vertices of the meshes. This is in fact a well-known finite element space, called the Hermite quintic or Argyris space, developed for solving 4th order partial differential equations (for which the inclusion in $H^2(\Omega)$ and therefore C^1 continuity is required). The shape functions are $\mathbb{P}_5(T)$ and the 21 degrees of freedom are the values of the function and all its first and second partial derivatives at the vertices and the integrals of the normal derivatives along edges. We then have a *discrete elasticity complex*

$$\mathbb{P}_1 \hookrightarrow Y_h \xrightarrow{J} S_h \xrightarrow{\operatorname{div}} V_h \to 0,$$

or, diagrammatically,



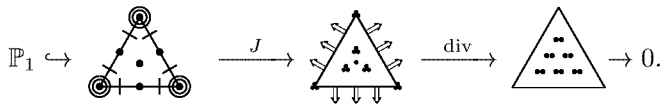
Moreover this sequence is exact and is coupled to the two-dimensional elasticity sequence via a commuting diagram:

$$\begin{array}{ccccccc}
\mathbb{P}_1 \hookrightarrow C^\infty(\Omega) & \xrightarrow{J} & C^\infty(\Omega,\mathbb{S}) & \xrightarrow{\text{div}} & C^\infty(\Omega,\mathbb{R}^3) & \to 0 \\
\downarrow \Pi_h^Y & & \downarrow \Pi_h^S & & \downarrow \Pi_h^V & \\
\mathbb{P}_1 \hookrightarrow Y_h & \xrightarrow{J} & S_h & \xrightarrow{\text{div}} & V_h & \to 0
\end{array}$$

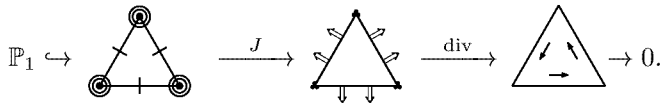
The right half of this diagram encodes the information necessary to establish the stability of our mixed finite element method.

The Hermite quintic finite elements arose naturally from our mixed finite elements to complete the commutative diagram. Had they not been long known, we could have used this procedure to devise a finite element space contained in $H^2(\Omega)$. In fact, on close scrutiny we can see that any stable mixed finite elements for elasticity with polynomial shape functions will give rise to a finite element space with polynomial shape functions contained in $H^2(\Omega)$. However, it is known that such spaces are difficult to construct and complicated. In fact, it can be proved that an H^2 finite element space must utilize shape functions of degree at least 5 and the first and second partial derivatives at the vertices must be among the degrees of freedom [14]. This helps explain why mixed finite elements for elasticity have proven so hard to devise. In particular, we can rigorously establish the stress elements must involve polynomials of degree 3, and that vertex degrees of freedom are unavoidable.

In addition to the element just described, elements of all greater orders are also introduced in [1]. The elements of next higher order can be seen as the final two elements in this discrete elasticity complex.

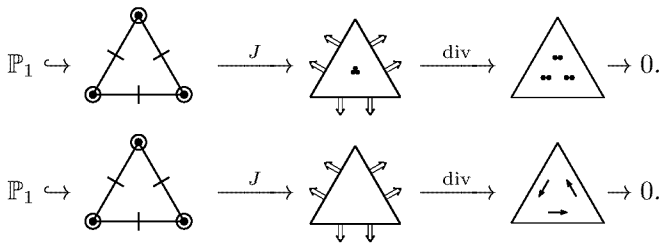


It is also possible to simplify the lowest order element slightly. To do this we reduce the displacement space from piecewise linear vector fields to piecewise rigid motions, and we replace the stress space with the inverse image under the divergence of the reduced displacement space. This leads to a stable element shown in this exact sequence:



Because of the unavoidable complexity of H^2 finite elements, practitioners solving 4th order equations often resort to *nonconforming* finite element approximations of H^2. This means that the finite element space does not belong to H^2

in that the function or the normal derivative may jump across element boundaries, but the spaces are designed so that jumps are small enough in some sense (e.g., on average). The error analysis is more complicated for nonconforming elements, since in addition to stability and approximation properties of the finite element space, one must analyze the *consistency error* arising from the jumps in the finite elements. In [2] Winther and the author investigated the the possibility of nonconforming mixed finite elements for elasticity, which, however are stable and convergent, and developed two such elements. These are related to nonconforming H^2 elements via nonconforming discrete elasticity complexes, two of which are pictured here:



In both cases the shape function space for the stress is contained between $\mathbb{P}_1(T, \mathbb{S})$ and $\mathbb{P}_2(T, \mathbb{S})$. The nonconforming H^2 finite element depicted in these diagrams was developed for certain 4th order problems in [11]. Note the nonconforming mixed elasticity elements are significantly simpler than the conforming ones (and, in particular, don't require vertex degrees of freedom).

References

[1] D. N. Arnold & R. Winther, Mixed finite elements for elasticity, *Numer. Math.*, 92(2001), 401–419.

[2] D. N. Arnold & R. Winther, Nonconforming mixed finite elements for elasticity, *Math. Models Methods Appl. Sci.*, to appear.

[3] I. Babuška & J. Osborn, Eigenvalue Problems, in: *Handbook of Numerical Analysis*, vol. II, P. G. Ciarlet & J. L. Lions, eds., Elsevier, 1991, 641–788.

[4] D. Boffi, P. Fernandes, L. Gastaldi & I. Perugia, Computational models of electromagnetic resonators: analysis of edge element approximation, *SIAM J. Numer. Anal.*, 36 (1999), 1264–1290.

[5] A. Bossavit, Whitney forms: a class of finite elements for three-dimensional computations in electromagnetism, *IEEE Proc. A*, 135 (1988), 493–500.

[6] F. Brezzi, On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers, *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.*, 8 (1974), 129–151.

[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, 1978.

[8] M. Eastwood, A complex from linear elasticity, *Rend. Circ. Mat. Palermo (2) Suppl.*, 63 (2000), 23–29.

[9] R. Hiptmair, Finite elements in computational electromagnetism, *Acta Numerica*, 11 (2002), 237–340.

[10] J.-C. Nedelec, Mixed finite elements in $\mathbb{R}^3$, *Numer. Math.*, 50 (1980), 315–341.
[11] T. K. Nilsen, X.-C. Tai & R. Winther, A robust nonconforming H^2-element, *Math. Comp.*, 70 (2001), 489–505.
[12] P. A. Raviart & J. M. Thomas, A mixed finite element method for second order elliptic problems, Springer Lecture Notes in Mathematics vol. 606, Springer-Verlag, 1977, 292–315.
[13] H. Whitney, *Geometric Integration Theory*, Princeton University Press, 1957.
[14] A. Ženišek, A general theorem on triangular $C^{(m)}$ finite elements, *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.*, 8 (1974), 119–127.

ICM 2002 · Vol. I · 159–178

Hyperbolic Systems of Conservation Laws in One Space Dimension

Alberto Bressan*

Abstract

Aim of this paper is to review some basic ideas and recent developments in the theory of strictly hyperbolic systems of conservation laws in one space dimension. The main focus will be on the uniqueness and stability of entropy weak solutions and on the convergence of vanishing viscosity approximations.

2000 Mathematics Subject Classification: 35L60, 35L65.

Keywords and Phrases: Hyperbolic system of conservation laws, Entropy weak solution, Vanishing viscosity.

1. Introduction

By a *system of conservation laws* in m space dimensions we mean a first order system of partial differential equations in divergence form:

$$\frac{\partial}{\partial t}U + \sum_{\alpha=1}^{m} \frac{\partial}{\partial x_\alpha} F_\alpha(U) = 0\,, \qquad\qquad U \in \mathbb{R}^n, \quad (t,x) \in \mathbb{R} \times \mathbb{R}^m.$$

The components of the vector $U = (U_1, \ldots, U_n)$ are the *conserved quantities.* Systems of this type express the balance equations of continuum physics, when small dissipation effects are neglected. A basic example is provided by the equations of non-viscous gases, accounting for the conservation of mass, momentum and energy. The subject is thus very classical, having a long tradition which can be traced back to Euler (1755) and includes contributions by Stokes, Riemann, Weyl and Von Neumann, among several others. The continued attention of analysts and mathematical physicists during the span of over two centuries, however, has not accounted for a comprehensive mathematical theory. On the contrary, as remarked in [Lx2], [D2], [S2], the field is still replenished with challenging open problems. In several space dimensions, not even the global existence of solutions is presently known, in any

* S.I.S.S.A., Via Beirut 4, Trieste 34014, Italy. E-mail: bressan@sissa.it

significant degree of generality. Until now, most of the analysis has been concerned with the one-dimensional case, and it is only here that basic questions could be settled. In the remainder of this paper we shall thus consider systems in one space dimension, referring to the books of Majda [M], Serre [S1] or Dafermos [D3] for a discussion of the multidimensional case.

Toward a rigorous mathematical analysis of solutions, the main difficulty that one encounters is the lack of regularity. Due to the strong nonlinearity of the equations and the absence of diffusion terms with smoothing effect, solutions which are initially smooth may become discontinuous within finite time. In the presence of discontinuities, most of the classical tools of differential calculus do not apply. Moreover, for general $n \times n$ systems, the powerful techniques of functional analysis cannot be used. In particular, solutions cannot be represented as fixed points of a nonlinear transformation, or in variational form as critical points of a suitable functional. Dealing with vector valued functions, comparison arguments based on upper and lower solutions do not apply either. Up to now, the theory of conservation laws has progressed largely by *ad hoc* methods. A survey of these techniques is the object of the present paper.

The Cauchy problem for a system of conservation laws in one space dimension takes the form

$$u_t + f(u)_x = 0, \tag{1.1}$$

$$u(0,x) = \bar{u}(x). \tag{1.2}$$

Here $u = (u_1, \ldots, u_n)$ is the vector of *conserved quantities*, while the components of $f = (f_1, \ldots, f_n)$ are the *fluxes*. We shall always assume that the flux function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is smooth and that the system is *strictly hyperbolic*, i. e., at each point u the Jacobian matrix $A(u) = Df(u)$ has n real, distinct eigenvalues

$$\lambda_1(u) < \cdots < \lambda_n(u). \tag{1.3}$$

As already mentioned, a distinguished feature of nonlinear hyperbolic systems is the possible loss of regularity. Even with smooth initial data, it is well known that the solution can develop shocks in finite time. Therefore, solutions defined globally in time can only be found within a space of discontinuous functions. The equation (1.1) must then be interpreted in distributional sense. A vector valued function $u = u(t,x)$ is a *weak solution* of (1.1) if

$$\iint \big[u\,\phi_t + f(u)\,\phi_x\big]\,dxdt = 0 \tag{1.4}$$

for every test function $\phi \in \mathcal{C}^1_c$, continuously differentiable with compact support. In particular, the piecewise constant function

$$u(t,x) \doteq \begin{cases} u^- & \text{if} \quad x < \lambda t\,, \\ u^+ & \text{if} \quad x > \lambda t\,, \end{cases} \tag{1.5}$$

is a weak solution of (1.1) if and only if the left and right states u^-, u^+ and the speed λ satisfy the famous Rankine-Hugoniot equations

$$f(u^+) - f(u^-) = \lambda\,(u^+ - u^-)\,. \tag{1.6}$$

When discontinuities are present, the weak solution of a Cauchy problem may not be unique. To single out a unique "good" solution, additional *entropy conditions* are usually imposed along shocks [Lx1], [L3]. These conditions often have a physical motivation, characterizing those solutions which can be recovered from higher order models, letting the diffusion or dispersion coefficients approach zero (see [D3]).

In one space dimension, the mathematical theory of hyperbolic systems of conservation laws has developed along two main lines.

1. The BV setting, pioneered by Glimm (1965). Solutions are here constructed within a space of functions with bounded variation, controlling the BV norm by a wave interaction potential.

2. The $\mathbf{L}^\infty$ setting, introduced by DiPerna (1983), based on weak convergence and a compensated compactness argument.

Both approaches yield results on the global existence of weak solutions. However, it is only in the BV setting that the well posedness of the Cauchy problem could recently be proved, as well as the stability and convergence of vanishing viscosity approximations. On the other hand, a counterexample in [BS] indicates that similar results cannot be expected, in general, for solutions in $\mathbf{L}^\infty$. In the remainder of this paper we thus concentrate on the theory of BV solutions, referring to [DP2] or [S1] for the alternative approach based on compensated compactness.

We shall first review the main ideas involved in the construction of weak solutions, based on the Riemann problem and the wave interaction functional. We then present more recent results on stability, uniqueness and characterization of entropy weak solutions. All this material can be found in the monograph [B3]. The last section contains an outline of the latest work on stability and convergence of vanishing viscosity approximations.

2. Existence of weak solutions

Toward the construction of more general solutions of (1.1), the basic building block is the *Riemann problem*, i.e. the initial value problem where the data are piecewise constant, with a single jump at the origin:

$$u(0,x) = \begin{cases} u^- & \text{if} \quad x < 0\,, \\ u^+ & \text{if} \quad x > 0\,. \end{cases} \tag{2.1}$$

Assuming that the amplitude $|u^+ - u^-|$ of the jump is small, this problem was solved in a classical paper of Lax [Lx1], under the additional hypothesis

(H) For each $i = 1, \ldots, n$, the i-th field is either *genuinely nonlinear*, so that $D\lambda_i(u) \cdot r_i(u) > 0$ for all u, or *linearly degenerate*, with $D\lambda_i(u) \cdot r_i(u) = 0$ for all u.

The solution is self-similar: $u(t,x) = U(x/t)$. It consists of $n+1$ constant states $\omega_0 = u^-$, $\omega_1, \ldots, \omega_n = u^+$ (see Fig. 1). Each couple of adiacent states ω_{i-1}, ω_i is separated either by a *shock* (the thick lines in Fig. 1) satisfying the Rankine Hugoniot equations, or else by a *centered rarefaction.* In this second case, the solution u varies continuously between ω_{i-1} and ω_i in a sector of the t-x-plane (the shaded region in Fig. 1) where the gradient u_x coincides with an i-eigenvector of the matrix $A(u)$.

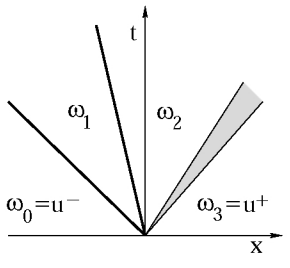


Figure 1

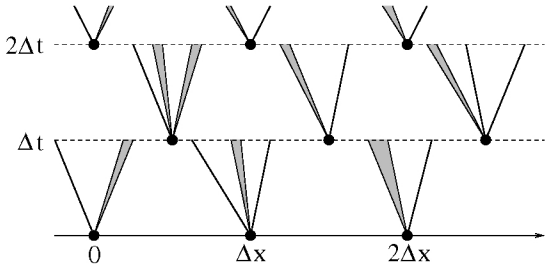


Figure 2

Approximate solutions to a more general Cauchy problem can be constructed by patching together several solutions of Riemann problems. In the Glimm scheme (Fig. 2), one works with a fixed grid in the x-t plane, with mesh sizes Δx, Δt. At time $t = 0$ the initial data is approximated by a piecewise constant function, with jumps at grid points. Solving the corresponding Riemann problems, a solution is constructed up to a time Δt sufficiently small so that waves generated by different Riemann problems do not interact. By a random sampling procedure, the solution $u(\Delta t, \cdot)$ is then approximated by a piecewise constant function having jumps only at grid points. Solving the new Riemann problems at every one of these points, one can prolong the solution to the next time interval $[\Delta t,\ 2\Delta t]$, etc...

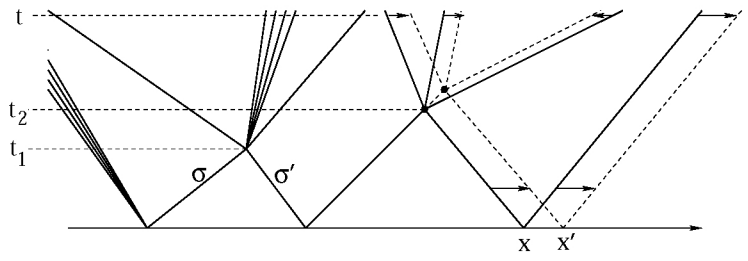


Figure 3

An alternative technique for contructing approximate solutions is by wave-

front tracking (Fig. 3). This method was introduced by Dafermos [D1] in the scalar case and later developed by various authors [DP1], [B1], [R], [BJ]. It now provides an efficient tool in the study of general $n \times n$ systems of conservation laws, both for theoretical and numerical purposes [B3], [HR].

The initial data is here approximated with a piecewise constant function, and each Riemann problem is solved approximately, within the class of piecewise constant functions. In particular, if the exact solution contains a centered rarefaction, this must be approximated by a *rarefaction fan*, containing several small jumps. At the first time t_1 where two fronts interact, the new Riemann problem is again approximately solved by a piecewise constant function. The solution is then prolonged up to the second interaction time t_2, where the new Riemann problem is solved, etc... The main difference is that in the Glimm scheme one specifies a priori the nodal points where the the Riemann problems are to be solved. On the other hand, in a solution constructed by wave-front tracking the locations of the jumps and of the interaction points depend on the solution itself, and no restarting procedure is needed.

In the end, both algorithms produce a sequence of approximate solutions, whose convergence relies on a compactness argument based on uniform bounds on the total variation. We sketch the main idea involved in these a priori BV bounds. Consider a piecewise constant function $u : \mathbb{R} \mapsto \mathbb{R}^n$, say with jumps at points $x_1 < x_2 < \cdots < x_N$. Call σ_α the amplitude of the jump at x_α. The *total strength of waves* is then defined as

$$V(u) \doteq \sum_\alpha |\sigma_\alpha|. \tag{2.2}$$

Clearly, this is an equivalent way to measure the total variation. Along a solution $u = u(t,x)$ constructed by front tracking, the quantity $V(t) = V\big(u(t,\cdot)\big)$ may well increase at interaction times. To provide global a priori bounds, following [G] one introduces a *wave interaction potential*, defined as

$$Q(u) = \sum_{(\alpha,\beta)\in\mathcal{A}} |\sigma_\alpha\, \sigma_\beta|, \tag{2.3}$$

where the summation runs over the set $\mathcal{A}$ of all couples of approaching waves. Roughly speaking, we say that two wave-fronts located at $x_\alpha < x_\beta$ are *approaching* if the one at x_α has a faster speed than the one at x_β (hence the two fronts are expected to collide at a future time). Now consider a time τ where two incoming wave-fronts interact, say with strengths σ, σ' (for example, take $\tau = t_1$ in Fig. 3). The difference between the outgoing waves emerging from the interaction and the two incoming waves σ, σ' is of magnitude $\mathcal{O}(1) \cdot |\sigma\sigma'|$. On the other hand, after time τ the two incoming waves are no longer approaching. This accounts for the decrease of the functional Q in (2.3) by the amount $|\sigma\sigma'|$. Observing that the new waves generated by the interaction could approach all other fronts, the change in the functionals V, Q across the interaction time τ is estimated as

$$\Delta V(\tau) = \mathcal{O}(1) \cdot |\sigma\sigma'|\,, \qquad\qquad \Delta Q(\tau) = -|\sigma\sigma'| + \mathcal{O}(1) \cdot |\sigma\sigma'|\, V(\tau-).$$

If the initial data has small total variation, for a suitable constant C_0 the quantity

$$\Upsilon(t) \doteq V\big(u(t,\cdot)\big) + C_0\, Q\big(u(t,\cdot)\big)$$

is monotone decreasing in time. This argument provides the uniform BV bounds on all approximate solutions. Using Helly's compactness theorem, one obtains the convergence of a subsequence of approximate solutions, and hence the global existence of a weak solution.

Theorem 1. *Let the system (1.1) be strictly hyperbolic and satisfy the assumptions (H). Then, for a sufficiently small $\delta > 0$ the following holds. For every initial condition $\bar u$ with*

$$\|\bar u\|_{\mathbf{L}^\infty} < \delta\,, \qquad\qquad \mathit{Tot.Var.}\{\bar u\} < \delta\,, \tag{2.4}$$

the Cauchy problem has a weak solution, defined for all times $t \geq 0$.

This result is based on careful analysis of solutions of the Riemann problem and on the use of a quadratic interaction functional (2.3) to control the creation of new waves. These techniques also provided the basis for subsequent investigations of Glimm and Lax [GL] and Liu [L2] on the asymptotic behavior of weak solutions as $t \to \infty$.

3. Stability

The previous existence result relied on a compactness argument which, by itself, does not provide informations on the uniqueness of solutions. A first understanding of the dependence of weak solutions on the initial data was provided by the analysis of front tracking approximations. The idea is to perturb the initial data by shifting the position of one of the jumps, say from x to a nearby point x' (see Fig. 3). By carefully estimating the corresponding shifts in the positions of all wave-fronts at a later time t, one obtains a bound on the $\mathbf{L}^1$ distance between the original and the perturbed approximate solution. After much technical work, this approach yielded a proof of the Lipschitz continuous dependence of solutions on the initial data, first in [BC1] for 2×2 systems, then in [BCP] for general $n \times n$ systems.

Theorem 2. *Let the system (1.1) be strictly hyperbolic and satisfy the assumptions (H). Then, for every initial data $\bar u$ satisfying (2.4) the weak solution obtained as limit of Glimm or front tracking approximations is unique and depends Lipschitz continuously on the initial data, in the $\mathbf{L}^1$ distance.*

These weak solutions can thus be written in the form $u(t,\cdot) = S_t\bar u$, as trajecories of a semigroup $S : \mathcal{D} \times [0,\infty[\mapsto \mathcal{D}$ on some domain $\mathcal{D}$ containing all functions with sufficiently small total variation. For some Lipschitz constants L, L' one has

$$\big\|S_t\bar u - S_s\bar v\big\|_{\mathbf{L}^1} \leq L\,\|\bar u - \bar v\|_{\mathbf{L}^1} + L'|t-s|\,, \tag{3.1}$$

for all $t, s \geq 0$ and initial data $\bar{u}, \bar{v} \in \mathcal{D}$.

An alternative proof of Theorem 2 was later achieved by a technique introduced by Liu and Yang in [LY] and presented in [BLY] in its final form. The heart of the matter is to construct a nonlinear functional, equivalent to the $\mathbf{L}^1$ distance, which is decreasing in time along every pair of solutions. We thus seek $\Phi = \Phi(u, v)$ and a constant C such that

$$\frac{1}{C} \cdot \|v - u\|_{\mathbf{L}^1} \leq \Phi(u, v) \leq C \cdot \|v - u\|_{\mathbf{L}^1}, \tag{3.2}$$

$$\frac{d}{dt}\Phi\big(u(t),\ v(t)\big) \leq 0. \tag{3.3}$$

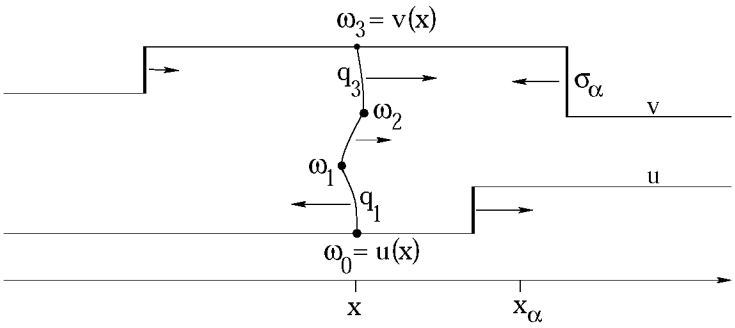


Figure 4

In connection with piecewise constant functions $u, v : \mathbb{R} \mapsto \mathbb{R}^n$ generated by a front tracking algorithm, this functional can be defined as follows (Fig. 4). At each point x, we connect the states $u(x)$, $v(x)$ by means of n shock curves. In other words, we construct intermediate states $\omega_0 = u(x), \omega_1, \ldots, \omega_n = v(x)$ such that each pair ω_{i-1}, ω_i is connected by an i-shock. These states can be uniquely determined by the implicit function theorem. Call $q_1, \ldots, q_n$, the strengths of these shocks. We regard $q_i(x)$ as the i-th scalar component of the jump $\big(u(x),\ v(x)\big)$. For some constant C', one clearly has

$$\frac{1}{C'} \cdot \big|v(x) - u(x)\big| \leq \sum_{i=1}^{n} \big|q_i(x)\big| \leq C' \cdot \big|v(x) - u(x)\big|. \tag{3.4}$$

The functional Φ is now defined as

$$\Phi(u, v) \doteq \sum_{i=1}^{n} \int_{-\infty}^{\infty} W_i(x)\, \big|q_i(x)\big|\, dx, \tag{3.5}$$

where the weights W_i take the form

$$\begin{aligned} W_i(x) \doteq 1 &+ \kappa_1 \cdot \big[\text{total strength of waves in } u \text{ and in } v \\ &\qquad \text{which approach the } i\text{-wave } q_i(x)\big] \\ &+ \kappa_2 \cdot \big[\text{wave interaction potentials of } u \text{ and of } v\big] \\ \doteq 1 &+ \kappa_1 V_i(x) + \kappa_2 \big[Q(u) + Q(v)\big] \end{aligned} \tag{3.6}$$

for suitable constants κ_1, κ_2. Notice that, by construction, $q_i(x)$ represents the strength of a fictitious shock wave located at x, travelling with a speed $\lambda_i(x)$ determined by the Rankine-Hugoniot equations. In (3.6), it is thus meaningful to consider the quantity

$$V_i(x) \doteq \sum_{\alpha \in \mathcal{A}_i(x)} |\sigma_\alpha| \,,$$

where the summation extends to all wave-fronts σ_α in u and in v which are *approaching* the i-shock $q_i(x)$. From (3.4) and the boundedness of the weights W_i, one easily derives (3.2). By careful estimates on the Riemann problem, one can prove that also (3.3) is approximately satisfied. In the end, by taking a limit of front tracking approximations, one obtains Theorem 2.

For general $n \times n$ systems, in (3.1) one finds a Lipschitz constant $L > 1$. Indeed, it is only in the scalar case that the semigroup is contractive and the theory of accretive operators and abstract evolution equations in Banach spaces can be applied, see [K], [C]. We refer to the flow generated by a system of conservation laws as a *Riemann semigroup*, because it is entirely determined by specifying how Riemann problems are solved. As proved in [B2], if two semigroups S, S' yield the same solutions to all Riemann problems, then they coincide, up to the choice of their domains.

From (3.1) one can deduce the error bound

$$\big\|w(T) - S_T w(0)\big\|_{\mathbf{L}^1} \leq L \cdot \int_0^T \left\{ \liminf_{h \to 0+} \frac{\big\|w(t+h) - S_h w(t)\big\|_{\mathbf{L}^1}}{h} \right\} dt \,, \tag{3.7}$$

valid for every Lipschitz continuous map $w : [0,T] \mapsto \mathcal{D}$ taking values inside the domain of the semigroup. We can think of $t \mapsto w(t)$ as an approximate solution of (1.1), while $t \mapsto S_t w(0)$ is the exact solution having the same initial data. According to (3.7), the distance at time T is bounded by the integral of an *instantaneous error rate*, amplified by the Lipschitz constant L of the semigroup.

Using (3.7), one can estimate the distance between a front tracking approximation and the corresponding exact solution. For approximate solutions constructed by the Glimm scheme, a direct application of this same formula is not possible because of the additional errors introduced by the restarting procedures at times $t_k \doteq k\,\Delta t$. However, relying on a careful analysis of Liu [L1], one can construct a front tracking approximate solution having the same initial and terminal values as the Glimm solution. By this technique, in [BM] the authors proved the estimate

$$\lim_{\Delta x \to 0} \frac{\big\|u^{\text{Glimm}}(T,\cdot) - u^{\text{exact}}(T,\cdot)\big\|_{\mathbf{L}^1}}{\sqrt{\Delta x} \cdot |\ln \Delta x|} = 0 \,. \tag{3.8}$$

In other words, letting the mesh sizes $\Delta x, \Delta t \to 0$ while keeping their ratio $\Delta x/\Delta t$ constant, the $\mathbf{L}^1$ norm of the error in the Glimm approximate solution tends to zero at a rate slightly slower than $\sqrt{\Delta x}$.

4. Uniqueness

The uniqueness and stability results stated in Theorem 2 refer to a special class of weak solutions: those obtained as limits of Glimm or front tracking approximations. For several applications, it is desirable to have a uniqueness theorem valid for general weak solutions, without reference to any particular constructive procedure. Results in this direction were proved in [BLF], [BG], [BLe]. They are all based on the error formula (3.7). In the proofs, one considers a weak solution $u = u(t,x)$ of the Cauchy problem (1.1)–(1.2). Assuming that u satisfies suitable entropy and regularity conditions, one shows that

$$\liminf_{h\to 0+} \frac{\|u(t+h) - S_h u(t)\|_{\mathbf{L}^1}}{h} = 0 \tag{4.1}$$

at almost every time t. By (3.7), u thus coincides with the semigroup trajectory $t \mapsto S_t u(0) = S_t \bar{u}$. Of course, this implies uniqueness. As an example, we state below the result of [BLe]. Consider the following assumptions:

(A1) (Conservation Equations) The function $u = u(t,x)$ is a weak solution of the Cauchy problem (1.1)–(1.2), taking values within the domain $\mathcal{D}$ of the semigroup S. More precisely, $u : [0,T] \mapsto \mathcal{D}$ is continuous w.r.t. the $\mathbf{L}^1$ distance. The initial condition (1.2) holds, together with

$$\iint \big[u\,\phi_t + f(u)\,\phi_x\big]\,dxdt = 0$$

for every $\mathcal{C}^1$ function ϕ with compact support contained inside the open strip $]0,T[\,\times\mathbb{R}$.

(A2) (Lax Entropy Condition) Let u have an approximate jump discontinuity at some point $(\tau,\xi) \in]0,T[\times\mathbb{R}$. In other words, assume that there exists states $u^-, u^+ \in \Omega$ and a speed $\lambda \in \mathbb{R}$ such that, calling

$$U(t,x) \doteq \begin{cases} u^- & \text{if} \quad x < \xi + \lambda(t-\tau), \\ u^+ & \text{if} \quad x > \xi + \lambda(t-\tau), \end{cases} \tag{4.2}$$

there holds

$$\lim_{\rho\to 0+} \frac{1}{\rho^2} \int_{\tau-\rho}^{\tau+\rho} \int_{\xi-\rho}^{\xi+\rho} \Big|u(t,x) - U(t,x)\Big|\,dxdt = 0. \tag{4.3}$$

Then, for some $i \in \{1,\ldots,n\}$, one has the entropy inequality:

$$\lambda_i(u^-) \geq \lambda \geq \lambda_i(u^+). \tag{4.4}$$

(A3) (Bounded Variation Condition) The function $x \mapsto u\big(\tau(x),x\big)$ has bounded variation along every Lipschitz continuous space-like curve $\big\{t = \tau(x)\big\}$, which satisfies $|d\tau/dx| < \delta$ a.e., for some constant $\delta > 0$ small enough.

Theorem 3. *Let $u = u(t,x)$ be a weak solution of the Cauchy problem (1.1)–(1.2) satisfying the assumptions (A1), (A2) and (A3). Then*

$$u(t,\cdot) = S_t\bar{u} \tag{4.5}$$

for all t. In particular, the solution that satisfies the three above conditions is unique.

An additional characterization of these unique solutions, based on local integral estimates, was given in [B2]. The underlying idea is as follows. In a forward neighborhood of a point (τ,ξ) where u has a jump, the weak solution u behaves much in the same way as the solution of the corresponding Riemann problem. On the other hand, on a region where its total variation is small, our solution u can be accurately approximated by the solution of a linear hyperbolic system with constant coefficients.

To state the result more precisely, we introduce some notations. Given a function $u = u(t,x)$ and a point (τ,ξ), we denote by $U^\sharp_{(u;\tau,\xi)}$ the solution of the Riemann problem with initial data

$$u^- = \lim_{x\to\xi-} u(\tau,x), \qquad\qquad u^+ = \lim_{x\to\xi+} u(\tau,x). \tag{4.6}$$

In addition, we define $U^\flat_{(u;\tau,\xi)}$ as the solution of the linear hyperbolic Cauchy problem with constant coefficients

$$w_t + \widehat{A}w_x = 0, \qquad\qquad w(0,x) = u(\tau,x). \tag{4.7}$$

Here $\widehat{A} \doteq A\big(u(\tau,\xi)\big)$. Observe that (4.7) is obtained from the quasilinear system

$$u_t + A(u)u_x = 0 \qquad\qquad (A = Df) \tag{4.8}$$

by "freezing" the coefficients of the matrix $A(u)$ at the point (τ,ξ) and choosing $u(\tau)$ as initial data. A new notion of "good solution" can now be introduced, by locally comparing a function u with the self-similar solution of a Riemann problem and with the solution of a linear hyperbolic system with constant coefficients. More precisely, we say that a function $u = u(t,x)$ is a **viscosity solution** of the system (1.1) if $t \mapsto u(t,\cdot)$ is continuous as a map with values into $\mathbf{L}^1_{\rm loc}$, and moreover the following integral estimates hold.

(i) At every point (τ,ξ), for every $\beta' > 0$ one has

$$\lim_{h\to 0+} \frac{1}{h} \int_{\xi-\beta' h}^{\xi+\beta' h} \Big| u(\tau+h,\ x) - U^\sharp_{(u;\tau,\xi)}(h,\ x-\xi) \Big|\, dx \ =\ 0. \tag{4.9}$$

(ii) There exist constants $C, \beta > 0$ such that, for every $\tau \geq 0$ and $a < \xi < b$, one has

$$\limsup_{h\to 0+} \frac{1}{h} \int_{a+\beta h}^{b-\beta h} \Big| u(\tau+h,\ x) - U^\flat_{(u;\tau,\xi)}(h,x) \Big|\, dx \leq C \cdot \Big(\text{Tot.Var.}\big\{u(\tau);\ ]a,\ b[\ \big\} \Big)^2. \tag{4.10}$$

As proved in [B2], this concept of viscosity solution completely characterizes semigroup trajectories.

Theorem 4. *Let $S : \mathcal{D} \times [0,\infty[\times \mathcal{D}$ be a semigroup generated by the system of conservation laws (1.1). A function $u : [0,T] \mapsto \mathcal{D}$ is a viscosity solution of (1.1) if and only if $u(t) = S_t u(0)$ for all $t \in [0,T]$.*

5. Vanishing viscosity approximations

A natural conjecture is that the entropic solutions of the hyperbolic system (1.1) actually coincide with the limits of solutions to the parabolic system

$$u^\varepsilon_t + f(u^\varepsilon)_x = \varepsilon\, u^\varepsilon_{xx}\,, \tag{5.1}$$

letting the viscosity coefficient $\varepsilon \to 0$. In view of the previous uniqueness results, one expects that the vanishing viscosity limit should single out the unique "good" solution of the Cauchy problem, satisfying the appropriate entropy conditions. In earlier literature, results in this direction were based on three main techniques:

1 - Comparison principles for parabolic equations. For a *scalar* conservation law, the existence, uniqueness and global stability of vanishing viscosity solutions was first established by Oleinik [O] in one space dimension. The famous paper by Kruzhkov [K] covers the more general class of $\mathbf{L}^\infty$ solutions and is also valid in several space dimensions.

2 - Singular perturbations. Let u be a piecewise smooth solution of the $n \times n$ system (1.1), with finitely many non-interacting, entropy admissible shocks. In this special case, using a singular perturbation technique, Goodman and Xin [GX] constructed a family of solutions u^ε to (5.1), with $u^\varepsilon \to u$ as $\varepsilon \to 0$.

3 - Compensated compactness. If, instead of a BV bound, only a uniform bound on the $\mathbf{L}^\infty$ norm of solutions of (5.1) is available, one can still construct a weakly convergent subsequence $u^\varepsilon \rightharpoonup u$. In general, we cannot expect that this weak limit satisfies the nonlinear equations (1.1). However, for a class of 2×2 systems, in [DP2] DiPerna showed that this limit u is indeed a weak solution of (1.1). The proof relies on a compensated compactness argument, based on the representation of the weak limit in terms of Young measures, which must reduce to a Dirac mass due to the presence of a large family of entropies.

Since the main existence and uniqueness results for hyperbolic systems of conservation laws are valid within the space of BV functions, it is natural to seek uniform BV bounds also for the viscous approximations u^ε in (5.1). This is indeed the main goal accomplished in [BB]. As soon as these BV bounds are established, the existence of a vanishing viscosity limit follows by a standard compactness argument. The uniqueness of the limit can then be deduced from the uniqueness theorem in [BG]. By further analysis, one can also prove the continuous dependence on the

initial data for the viscous approximations u^ε, in the $\mathbf{L}^1$ norm. Remarkably, these results are valid for general $n \times n$ strictly hyperbolic systems, not necessarily in conservation form.

Theorem 5. *Consider the Cauchy problem for a strictly hyperbolic system with viscosity*

$$u^\varepsilon_t + A(u^\varepsilon)u^\varepsilon_x = \varepsilon\, u^\varepsilon_{xx}, \qquad u^\varepsilon(0,x) = \bar u(x)\,. \tag{5.2}$$

Then there exist constants C, L, L' and $\delta > 0$ such that the following holds. If

$$Tot.Var.\{\bar u\} < \delta\,, \qquad \|\bar u(x)\|_{\mathbf{L}^\infty} < \delta\,, \tag{5.3}$$

then for each $\varepsilon > 0$ the Cauchy problem (5.2) has a unique solution u^ε, defined for all $t \geq 0$. Adopting a semigroup notation, this will be written as $t \mapsto u^\varepsilon(t,\cdot) \doteq S^\varepsilon_t \bar u$. In addition, one has:

BV bounds :
$$Tot.Var.\big\{S^\varepsilon_t \bar u\big\} \leq C\; Tot.Var.\{\bar u\}\,. \tag{5.4}$$

$\mathbf{L}^1$ stability :
$$\big\|S^\varepsilon_t \bar u - S^\varepsilon_t \bar v\big\|_{\mathbf{L}^1} \leq L\,\big\|\bar u - \bar v\big\|_{\mathbf{L}^1}\,, \tag{5.5}$$

$$\big\|S^\varepsilon_t \bar u - S^\varepsilon_s \bar u\big\|_{\mathbf{L}^1} \leq L'\left(|t-s| + \big|\sqrt{\varepsilon t} - \sqrt{\varepsilon s}\,\big|\right). \tag{5.6}$$

Convergence. *As $\varepsilon \to 0+$, the solutions u^ε converge to the trajectories of a semigroup S such that*

$$\big\|S_t\bar u - S_s \bar v\big\|_{\mathbf{L}^1} \leq L\,\|\bar u - \bar v\|_{\mathbf{L}^1} + L'\,|t-s|\,. \tag{5.7}$$

These vanishing viscosity limits can be regarded as the unique **vanishing viscosity solutions** *of the hyperbolic Cauchy problems*

$$u_t + A(u)u_x = 0, \qquad u(0,x) = \bar u(x)\,. \tag{5.8}$$

In the conservative case where $A(u) = Df(u)$ for some flux function f, the vanishing viscosity solution is a weak solution of

$$u_t + f(u)_x = 0, \qquad u(0,x) = \bar u(x)\,, \tag{5.9}$$

satisfying the Liu admissibility conditions [L3]. Moreover, the vanishing viscosity solutions are precisely the same as the viscosity solutions defined at (4.9)–(4.10) in terms of local integral estimates.

The key step in the proof is to establish a priori bounds on the total variation of solutions of

$$u_t + A(u)u_x = u_{xx} \tag{5.10}$$

uniformly valid for all times $t \in [0, \infty[$. We outline here the main ideas.

(i) At each point (t,x) we decompose the gradient along a suitable basis of unit vectors $\tilde{r}_i$, say

$$u_x = \sum v_i \tilde{r}_i \,. \tag{5.11}$$

(ii) We then derive an equation describing the evolution of these gradient components

$$v_{i,t} + (\tilde{\lambda}_i v_i)_x - v_{i,xx} = \phi_i \,. \tag{5.12}$$

(iii) Finally, we show that all source terms $\phi_i = \phi_i(t,x)$ are integrable. Hence, for all $\tau > 0$,

$$\big\|v_i(\tau,\cdot)\big\|_{\mathbf{L}^1} \leq \big\|v_i(0,\cdot)\big\|_{\mathbf{L}^1} + \int_0^\infty \int_{I\!R} \big|\phi_i(t,x)\big|\, dxdt < \infty \,. \tag{5.13}$$

In this connection, it seems natural to decompose the gradient u_x along the eigenvectors of the hyperbolic matrix $A(u)$. This approach however does NOT work. In the case where the solution u is a travelling viscous shock profile, we would obtain source terms which are not identically zero. Hence they are certainly not integrable over the domain $\big\{t>0\,, x \in I\!R\big\}$.

An alternative approach, proposed by S. Bianchini, is to decompose u_x as a *sum of gradients of viscous travelling waves.* By a viscous travelling i-wave we mean a solution of (5.10) having the form

$$w(t,x) = U(x-\sigma t)\,, \tag{5.14}$$

where the speed σ is close to the i-th eigenvalue λ_i of the hyperbolic matrix A. Clearly, the function U must provide a solution to the second order O.D.E.

$$U'' = \big(A(U) - \sigma\big)U'. \tag{5.15}$$

The underlying idea for the decomposition is as follows. At each point (t,x), given (u, u_x, u_{xx}), we seek travelling wave profiles $U_1, \ldots, U_n$ such that

$$U_i(x) = u(x), \qquad\qquad i = 1, \ldots, n\,, \tag{5.16}$$

$$\sum_i U_i'(x) = u_x(x)\,, \qquad\qquad \sum_i U_i''(x) = u_{xx}(x)\,. \tag{5.17}$$

In general, the system of algebraic equations (5.16)–(5.17) admits infinitely many solutions. A unique solution is singled out by considering only those travelling profiles U_i that lie on a suitable *center manifold* $\mathcal{M}_i$. We now call $\tilde{r}_i$ the unit vector parallel to U_i', so that $U_i' = v_i \tilde{r}_i$ for some scalar v_i. The decomposition (5.11) is then obtained from the first equation in (5.17).

Toward the BV estimate, the second part of the proof consists in deriving the equation (5.12) and estimating the integrals of the source terms ϕ_i. Here the main

idea is that these source terms can be regarded as generated by wave interactions. In analogy with the hyperbolic case considered by Glimm [G], the total amount of these interactions can be controlled by suitable Lyapunov functionals. We describe here the main ones.

1. Consider first two independent, scalar diffusion equations with strictly different drifts:

$$\begin{cases} z_t + \big[\lambda(t,x)z\big]_x - z_{xx} = 0\,, \\ z^*_t + \big[\lambda^*(t,x)z^*\big]_x - z^*_{xx} = 0\,, \end{cases}$$

assuming that

$$\inf_{t,x} \lambda^*(t,x) - \sup_{t,x} \lambda(t,x) \geq c > 0\,.$$

We regard z as the density of waves with a slow speed λ and z^* as the density of waves with a fast speed λ^*. A *transversal interaction potential* is defined as

$$Q(z,z^*) \doteq \frac{1}{c} \int\!\!\int_{\mathbb{R}^2} K(x_2 - x_1) \big|z(x_1)\big|\, \big|z^*(x_2)\big|\, dx_1 dx_2\,, \tag{5.18}$$

$$K(y) \doteq \begin{cases} e^{-cy/2} & \text{if} \quad y > 0\,, \\ 1 & \text{if} \quad y \leq 0\,. \end{cases} \tag{5.19}$$

One can show that this functional Q is monotonically decreasing along every couple of solutions z, z^*. The total amount of interaction between fast and slow waves can now be estimated as

$$\begin{aligned} \int_0^\infty \int_{\mathbb{R}} \big|z(t,x)\big|\, \big|z^*(t,x)\big|\, dxdt \;&\leq\; -\int_0^\infty \left[\frac{d}{dt} Q\big(z(t),\, z^*(t)\big)\right] dt \\ &\leq Q\big(z(0),\, z^*(0)\big) \;\leq\; \frac{1}{c} \int_{\mathbb{R}} \big|z(0,x)\big|\, dx \cdot \int_{\mathbb{R}} \big|z^*(0,x)\big|\, dx\,. \end{aligned}$$

By means of Lyapunov functionals of this type one can control all source terms in (5.12) due to the interaction of waves of different families.

2. To control the interactions between waves of the same family, we seek functionals which are decreasing along every solution of a scalar viscous conservation law

$$u_t + g(u)_x = u_{xx}\,. \tag{5.20}$$

For this purpose, to a scalar function $x \mapsto u(x)$ we associate the curve in the plane

$$\gamma \doteq \begin{pmatrix} u \\ g(u) - u_x \end{pmatrix} \;=\; \begin{pmatrix} \text{conserved quantity} \\ \text{flux} \end{pmatrix}. \tag{5.21}$$

In connection with a solution $u = u(t,x)$ of (5.20), the curve γ evolves according to

$$\gamma_t + g'(u)\gamma_x = \gamma_{xx}\,. \tag{5.22}$$

Notice that the vector $g'(u)\gamma_x$ is parallel to γ, hence the presence of this term in (5.22) only amounts to a reparametrization of the curve, and does not affect its shape. The curve thus evolves in the direction of curvature. An obvious Lyapunov functional is the *length* of the curve. In terms of the variables

$$\gamma_x = \begin{pmatrix} v \\ w \end{pmatrix} \doteq \begin{pmatrix} u_x \\ -u_t \end{pmatrix}, \tag{5.23}$$

this length is given by

$$L(\gamma) \doteq \int |\gamma_x|\, dx = \int \sqrt{v^2 + w^2}\, dx\,. \tag{5.24}$$

We can estimate the rate of decrease in the length as

$$-\frac{d}{dt}L\big(\gamma(t)\big) \;=\; \int_{I\!R} \frac{|v|\,\big[(w/v)_x\big]^2}{\big(1+(w/v)^2\big)^{3/2}}\, dx \;\geq\; \frac{1}{(1+\delta^2)^{3/2}} \int_{|w/v|\leq\delta} |v|\,\big[(w/v)_x\big]^2\, dx\,, \tag{5.25}$$

for any given constant $\delta > 0$. This yields a useful a priori estimate on the integral on the right hand side of (5.25).

3. In connection with the same curve γ in (5.21), we now introduce another functional, defined in terms of a wedge product.

$$Q(\gamma) \doteq \frac{1}{2}\iint_{x<x'} \big|\gamma_x(x) \wedge \gamma_x(x')\big|\, dx\, dx'\,. \tag{5.26}$$

For any curve that moves in the plane in the direction of curvature, one can show that this functional is monotone decreasing and its decrease bounds the area swept by the curve: $|dA| \leq -dQ$.

Using (5.22)–(5.23) we now compute

$$-\frac{dQ}{dt} \geq \left|\frac{dA}{dt}\right| = \int |\gamma_t \wedge \gamma_x|\, dx = \int |\gamma_{xx} \wedge \gamma_x|\, dx = \int |v_x w - v w_x|\, dx\,.$$

Integrating w.r.t. time, we thus obtain another useful a priori bound:

$$\int_0^\infty \int |v_x w - v w_x|\, dx\, dt \;\leq\; \int_0^\infty \left|\frac{dQ\big(\gamma(t)\big)}{dt}\right| dt \;\leq\; Q\big(\gamma(0)\big)\,.$$

Together, the functionals in (5.24) and (5.26) allow us to estimate all source terms in (5.12) due to the interaction of waves of the same family.

This yields the $\mathbf{L}^1$ estimates on the source terms ϕ_i, in (5.12), proving the uniform bounds on the total variation of a solution u of (5.10). See [BB] for details.

Next, to prove the uniform stability of all solutions of the parabolic system (5.10) having small total variation, we consider the linearized system describing the

evolution of a first order variation. Inserting the formal expansion $u = u_0 + \epsilon z + O(\epsilon^2)$ in (5.10), we obtain

$$z_t + \big[DA(u) \cdot z\big] u_x + A(u) z_x = z_{xx} \,. \tag{5.27}$$

Our basic goal is to prove the bound

$$\big\|z(t)\big\|_{\mathbf{L}^1} \leq L \,\big\|z(0)\big\|_{\mathbf{L}^1} \,, \tag{5.28}$$

for some constant L and all $t \geq 0$ and every solution z of (5.27). By a standard homotopy argument, from (5.28) one easily deduces the Lipschitz continuity of the solution of (5.8) on the initial data. Namely, for every couple of solutions $u, \tilde{u}$ with small total variation one has

$$\big\|u(t) - \tilde{u}(t)\big\|_{\mathbf{L}^1} \leq L \,\big\|u(0) - \tilde{u}(0)\big\|_{\mathbf{L}^1} \,. \tag{5.29}$$

To prove (5.28) we decompose the vector z as a sum of scalar components: $z = \sum_i h_i \tilde{r}_i$, write an evolution equation for these components:

$$h_{i,t} + (\tilde{\lambda}_i h_i)_x - h_{i,xx} = \hat{\phi}_i \,,$$

and show that the source terms $\hat{\phi}_i$ are integrable on the domain $\{t > 0\,, x \in I\!R\}$.

For every initial data $u(0, \cdot) = \bar{u}$ with small total variation, the previous arguments yield the existence of a unique global solution to the parabolic system (5.8), depending Lipschitz continuously on the initial data, in the $\mathbf{L}^1$ norm. Performing the rescaling $t \mapsto t/\varepsilon$, $x \mapsto x/\varepsilon$, we immediately obtain the same results for the Cauchy problem (5.2). Adopting a semigroup notation, this solution can be written as $u^\varepsilon(t, \cdot) = S^\varepsilon_t \bar{u}$. Thanks to the uniform bounds on the total variation, a compactness argument yields the existence of a strong limit in $\mathbf{L}^1_{\rm loc}$

$$u = \lim_{\varepsilon_m \to 0} u^{\varepsilon_m} \tag{5.30}$$

at least for some subsequence $\varepsilon_m \to 0$. Since the u^ε depend continuously on the initial data, with a uniform Lipschitz constant, the same is true of the limit solution $u(t, \cdot) = S_t \bar{u}$. In the conservative case where $A(u) = Df(u)$, it is not difficult to show that this limit u actually provides a weak solution to the Cauchy problem (1.1)–(1.2).

The only remaining issue is to show that the limit in (5.30) is unique, i.e. it does not depend on the subsequence $\{\varepsilon_m\}$. In the standard conservative case, this fact can already be deduced from the uniqueness result in [BG]. In the general case, uniqueness is proved in two steps. First we show that, in the special case of a Riemann problem, the solution obtained as vanishing viscosity limit is unique and can be completely characterized. To conclude the proof, we then rely on the same

general argument as in [B2]: if two Lipschitz semigroups S, S' provide the same solutions to all Riemann problems, then they must coincide. See [BB] for details.

6. Concluding remarks

1. A classical tool in the analysis of first order hyperbolic systems is the *method of characteristics.* To study the system

$$u_t + A(u)u_x = 0\,,$$

one decomposes the solution along the eigenspaces of the matrix $A(u)$. The evolution of these components is then described by a family of O.D.E's along the *characteristic curves.* In the t-x plane, these are the curves which satisfy $dx/dt = \lambda_i\big(u(t,x)\big)$. The local decomposition (5.16)–(5.17) in terms of viscous travelling waves makes it possible to implement this "hyperbolic" approach also in connection with the parabolic system (5.10). In this case, the projections are taken along the vectors $\tilde{r}_i$, while the characteristic curves are defined as $dx/dt = \sigma_i$, where σ_i is the speed of the i-th travelling wave. Notice that in the hyperbolic case the projections and the wave speeds depend only on the state u, through the eigenvectors $r_i(u)$ and the eigenvalues $\lambda_i(u)$ of the matrix $A(u)$. On the other hand, in the parabolic case the construction involves the derivatives u_x, u_{xx} as well.

2. In nearly all previous works on BV solutions for systems of conservation laws, following [G] the basic estimates on the total variation were obtained by a careful study of the Riemann problem and of elementary wave interactions. The Riemann problem also takes the center stage in all earlier proofs of the stability of solutions [BC1], [BCP], [BLY]. In this connection, the hypothesis (H) introduced by Lax [Lx1] is widely adopted in the literature. It guarantees that solutions of the Riemann problem have a simple structure, consisting of at most n elementary waves (shocks, centered rarefactions or contact discontinuities). If the assumption (H) is dropped, some results on global existence [L3], and continuous dependence [AM] are still available, but their proofs become far more technical. On the other hand, the approach introduced in [BB] marks the first time where uniform BV estimates are obtained without any reference to Riemann problems. Global existence and stability of weak solutions are obtained for the whole class of strictly hyperbolic systems, regardless of the hypothesis (H).

3. For the viscous system of conservation laws

$$u_t + f(u)_x = u_{xx}\,,$$

previous results in [L4], [SX], [SZ], [Yu] have established the stability of special types of solutions, for example travelling viscous shocks or viscous rarefactions. Taking $\varepsilon = 1$ in (5.2), from Theorem 5 we obtain the uniform Lipschitz stability (w..r.t. the $\mathbf{L}^1$ distance) of ALL viscous solutions with sufficiently small total variation. An

interesting alternative technique for proving stability of viscous solutions, based on spectral methods, was recently developed in [HZ].

4. In the present survey we only considered initial data with small total variation. This is a convenient setting, adopted in much of the current literature, which guarantees the global existence of BV solutions of (1.1) and captures the main features of the problem. A recent example constructed by Jenssen [J] shows that, for initial data with large total variation, the $\mathbf{L}^\infty$ norm of the solution can blow up in finite time. In this more general setting, one expects that the existence and uniqueness of weak solutions, together with the convergence of vanishing viscosity approximations, should hold locally in time as long as the total variation remains bounded. For the hyperbolic system (1.1), results on the local existence and stability of solutions with large BV data can be found in [Sc] and [BC2], respectively. Because of the counterexample in [BS], on the other hand, similar well posedness results are not expected in the general $\mathbf{L}^\infty$ case.

References

[AM] F. Ancona and A. Marson, Well posedness for general 2×2 systems of conservation laws, *Amer. Math. Soc. Memoir*, to appear.

[BaJ] P. Baiti and H. K. Jenssen, On the front tracking algorithm, *J. Math. Anal. Appl.* **217** (1998), 395–404.

[BB] S. Bianchini and A. Bressan, Vanishing viscosity solutions of nonlinear hyperbolic systems, preprint S.I.S.S.A., Trieste 2001.

[B1] A. Bressan, Global solutions to systems of conservation laws by wave-front tracking, *J. Math. Anal. Appl.* **170** (1992), 414–432.

[B2] A. Bressan, The unique limit of the Glimm scheme, *Arch. Rational Mech. Anal.* **130** (1995), 205–230.

[B3] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem.* Oxford University Press, 2000.

[BC1] A. Bressan and R. M. Colombo, The semigroup generated by 2×2 conservation laws, *Arch. Rational Mech. Anal.* **133** (1995), 1–75.

[BC2] A. Bressan and R. M. Colombo, Unique solutions of 2×2 conservation laws with large data, *Indiana Univ. Math. J.* **44** (1995), 677–725.

[BCP] A. Bressan, G. Crasta and B. Piccoli, Well posedness of the Cauchy problem for $n \times n$ conservation laws, *Amer. Math. Soc. Memoir* **694** (2000).

[BG] A. Bressan and P. Goatin, Oleinik type estimates and uniqueness for $n \times n$ conservation laws, *J. Diff. Equat.* **156** (1999), 26–49.

[BLF] A. Bressan and P. LeFloch, Uniqueness of weak solutions to systems of conservation laws, *Arch. Rat. Mech. Anal.* **140** (1997), 301–317.

[BLe] A. Bressan and M. Lewicka, A uniqueness condition for hyperbolic systems of conservation laws, *Discr. Cont. Dynam. Syst.* **6** (2000), 673–682.

[BLY] A. Bressan, T. P. Liu and T. Yang, L^1 stability estimates for $n \times n$ conservation laws, *Arch. Rational Mech. Anal.* **149** (1999), 1–22.

[BM] A. Bressan and A. Marson, Error bounds for a deterministic version of the Glimm scheme, *Arch. Rat. Mech. Anal.* **142** (1998), 155–176.

[BS] A. Bressan and W. Shen, Uniqueness for discontinuous O.D.E. and conservation laws, *Nonlinear Analysis, T. M. A.* **34** (1998), 637–652.

[C] M. Crandall, The semigroup approach to first-order quasilinear equations in several space variables, *Israel J. Math.* **12** (1972), 108–132.

[D1] C. Dafermos, Polygonal approximations of solutions of the initial value problem for a conservation law, *J. Math. Anal. Appl.* **38** (1972), 33–41.

[D2] C. Dafermos, Hyperbolic systems of conservation laws, *Proceedings of the International Congress of Mathematicians, Zürich 1994*, Birchäuser (1995), 1096–1107.

[D3] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin 2000.

[DP1] R. DiPerna, Global existence of solutions to nonlinear hyperbolic systems of conservation laws, *J. Diff. Equat.* **20** (1976), 187–212.

[DP2] R. DiPerna, Convergence of approximate solutions to conservation laws, *Arch. Rational Mech. Anal.* **82** (1983), 27–70.

[G] J. Glimm, Solutions in the large for nonlinear hyperbolic systems of equations, *Comm. Pure Appl. Math.* **18** (1965), 697–715.

[GL] J. Glimm and P. Lax, Decay of solutions of systems of nonlinear hyperbolic conservation laws, *Amer. Math. Soc. Memoir* **101** (1970).

[GX] J. Goodman and Z. Xin, Viscous limits for piecewise smooth solutions to systems of conservation laws, *Arch. Rational Mech. Anal.* **121** (1992), 235–265.

[HR] H. Holden and N. H. Risebro *Front Tracking for Hyperbolic Conservation Laws*, Springer Verlag, New York 2002.

[HZ] P. Howard and K. Zumbrun, Pointwise semigroup methods for stability of viscous shock waves, *Indiana Univ. Math. J.* **47** (1998), 727–841.

[K] S. Kruzhkov, First order quasilinear equations with several space variables, *Math. USSR Sbornik* **10** (1970), 217–243.

[J] H. K. Jenssen, Blowup for systems of conservation laws, *SIAM J. Math. Anal.* **31** (2000), 894–908.

[Lx1] P. Lax, Hyperbolic systems of conservation laws II, *Comm. Pure Appl. Math.* **10** (1957), 537–566.

[Lx2] P. Lax, Problems solved and unsolved concerning nonlinear P.D.E., *Proccedings of the International Congress of Mathematicians, Warszawa 1983.* Elsevier Science Pub. (1984), 119–138.

[L1] T. P. Liu, The deterministic version of the Glimm scheme, *Comm. Math. Phys.* **57** (1977), 135–148.

[L2] T. P. Liu, Linear and nonlinear large time behavior of solutions of general systems of hyperbolic conservation laws, *Comm. Pure Appl. Math.* **30** (1977), 767–796.

[L3] T. P. Liu, Admissible solutions of hyperbolic conservation laws, *Amer. Math. Soc. Memoir* **240** (1981).

[L4] T. P. Liu, Nonlinear stability of shock waves, *Amer. Math. Soc. Memoir* **328** (1986).

[LY] T. P. Liu and T. Yang, L^1 stability for 2×2 systems of hyperbolic conservation laws, *J. Amer. Math. Soc.* **12** (1999), 729–774.

[M] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, 1984.

[O] O. Oleinik, Discontinuous solutions of nonlinear differential equations (1957), *Amer. Math. Soc. Translations* **26**, 95–172.

[R] N. H. Risebro, A front-tracking alternative to the random choice method, *Proc. Amer. Math. Soc.* **117** (1993), 1125–1139.

[Sc] S. Schochet, Sufficient conditions for local existence via Glimm's scheme for large BV data, *J. Differential Equations* **89** (1991), 317–354.

[S1] D. Serre, *Systems of Conservation Laws I, II*, Cambridge University Press, 2000.

[S2] D. Serre, Systems of conservation laws : A challenge for the XXIst century, *Mathematics Unlimited - 2001 and beyond*, B. Engquist and W. Schmid eds., Springer-Verlag, 2001.

[SX] A. Szepessy and Z. Xin, Nonlinear stability abd viscous shocks, *Arch. Rational Mech. Anal.* **122** (1993), 53–103.

[SZ] A. Szepessy and K. Zumbrun, Stability of rarefaction waves in viscous media, *Arch. Rational Mech. Anal.* **133** (1996), 249–298.

[Yu] S. H. Yu, Zero-dissipation limit of solutions with shocks for systems of hyperbolic conservation laws, *Arch. Rational Mech. Anal.* **146** (1999), 275–370.

ICM 2002 · Vol. I · 179–187

Non Linear Elliptic Theory and the Monge-Ampere Equation

Luis A. Caffarelli*

Abstract

The Monge-Ampere equation, plays a central role in the theory of fully non linear equations. In fact we will like to show how the Monge-Ampere equation, links in some way the ideas comming from the calculus of variations and those of the theory of fully non linear equations.

2000 Mathematics Subject Classification: 35J15, 35J20, 35J70.

When learning complex analysis, it was a remarkable fact that the real part u of an analytic function, just because it satisfies the equation:

$$u_{xx} + u_{yy} = \Delta u = 0$$

(Laplace's equation) is real analytic, and furthermore, the oscillation of u in any given domain U, controls *all* the derivatives of u, of *any* order, in any subset $\bar{U}$, compactly contained in U.

One can give three, essentially different explanations of this phenomena.

a) Integral representations (Cauchy integral, for instance)**.** This gives rise to many of the modern aspects of real and harmonic analysis: fundamental solutions, singular integrals, pseudo-differential operators, etc. For our discussion, an important consequence of this theory are the Schauder and Calderon-Zygmund estimates.

Heuristically, they say that if we have a solution of an equation

$$A_{ij}(x) D_{ij} u = 0$$

and $A_{ij}(x)$ is, in a given functional space, a small perturbation of the Laplacian then $D_{ij}u$ is actually in the same functional space as A_{ij}. For instance, if $[A_{ij}]$ is Hölder continuous ($C^\alpha(\bar{U})$) and positive definite, we can transform it to the identity (the Laplacian) at any given point x_0 by an affine transformation, and will remain close to it in a neighborhood. Thus $D_{ij}u$ will also be $C^\alpha(\bar{U})$.

*Department of Mathematics, University of Texas at Austin, Austin, TX 78712, USA. E-mail: caffarel@math.utexas.edu

b) Energy considerations. Harmonic functions, u, are also local minimizers of the Dirichlet integral

$$E(v) = \int (\nabla v)^2 \, dx \ .$$

That is, if we change u to w, in $\bar{U} \subset\subset U$

$$E(w)|_{\bar{U}} \geq E(u)|_{\bar{U}} \ .$$

This gives rise to the theory of calculus of variations (minimal surface, harmonic maps, elasticity, fluid dynamics).

One is mainly concerned, there, with equations (or systems) of the form

$$D_i F_i(\nabla u, X) = 0 \ . \tag{1}$$

For instance, in the case in which u is a local minimizer of

$$E(u) = \int \mathcal{F}(\nabla u, X) \, dx$$

(1) is simply the Euler-Lagrange equation associated to E:

$$F_i = \nabla_p \mathcal{F} \ .$$

If we attempt to write (1) in second derivatives form, we get

$$F_{i,j}(\nabla u, X) D_{ij} u + \cdots = 0 \ .$$

This strongly suggests that in order for the variational problem to be "elliptic", like the Laplacian, $F_{i,j}$ should be positive definite, that is $\mathcal{F}$ should be strictly convex.

It also leads to the natural strategy of showing that ∇u, that in principle is only in L^2 (finite energy), is in fact Hölder continuous. Reaching this regularity allows us to apply the (linear) Schauder theory.

That implies $D_{ij} u$ is $C^\alpha(\bar{U})$, thus ∇u is $C^{1,\alpha}(\bar{U})$, and so on (the bootstrapping method).

The difficulty with this approach is that solutions, u, are invariant under R^{n+1}-dialations of their graphs.

This fact keeps the class of Lipschitz functions (bounded gradients) invariant. There is no reason, thus, to expect that this equation will "improve" under dialations. The fact that ∇u is indeed Hölder continuous is the celebrated De Giorgi's theorem, that solved the nineteenth Hilbert's problem:

De Giorgi looked at the equation that first derivatives, u_α satisfy

$$D_i F_{ij}(\nabla u) D_j u_\alpha = 0.$$

He thought of $F_{ij}(\nabla u)$ as elliptic coefficients $A_{ij}(x)$ that had no regularity whatsoever, and he proved that any solution w of

$$D_i A_{ij}(x) D_j w = 0$$

was Hölder continuous

$$\|w\|_{C^\alpha(\bar{U})} \leq C\|w\|_{L^2(U)} \ .$$

De Giorgi's theorem is in fact a linear one, but for a new invariant class of equations. No matter how the solution (and the equation) is renormalized, it stays far from the constant coefficient theory, and a radically new idea surfaces: if we have a class of functions for which at every scale, in some average sense, the function controls its derivatives (the energy inequality), further regularity follows.

Finally, the third approach is

c) Comparison principle. Two solutions u_1, u_2 of $\Delta u = 0$ cannot "touch without crossing". That is, if $u_1 - u_2$ is positive it cannot become zero in some interior point, X_0, of U.

Again, heuristically, this is because the function

$$F(D^2u) = \Delta u = \text{Trace}[D^2u]$$

is a monotone function of the Hessian matrix $[D_{ij}u]$ and, thus, in some sense, we must have $F(D^2u_1)$ ">" $F(D^2u_2)$ at X_0 (or nearby).

The natural family of equations to consider in this context, is then

$$F(D^2u) = 0$$

for F a strictly monotone function of D^2u.

Such type of equations appear in differential geometry. For instance, the coefficients of the characteristic polynomial of the Hessian

$$P(\lambda) = \det(D^2u - \lambda I)$$

are such equations if we restrict D^2u to stay in the appropriate set of $R^{n\times n}$. If λ_i denote the eigenvalues of D^2u

$$\begin{aligned} C_1 &= \Delta u = \sum \lambda_i \qquad \text{(Laplace)} \\ C_2 &= \sum_{i\neq j} \lambda_k\lambda_j \ldots \\ C_n &= \prod \lambda_i = \det D^2u \qquad \text{(Monge-Ampere)} \ . \end{aligned}$$

In the case of $C_n = \det D^2u = \prod \lambda_i$ is a monotone function of the Hessian provided that all λ_i's are positive. That is, provided that the function, u, under consideration is convex.

If $F(D^2u, X)$ is *uniformly elliptic*, that is, if F is strictly monotone as a function of the Hessian, or in differential form,

$$F_{ij}(M) = D_{m_{ij}}F$$

is uniformly positive definite, then solutions of $F(D^2u)$ are $C^{1,\alpha}(\bar{U})$. As in the divergence case, this is because first derivatives u_α satisfy an elliptic operator,

$$F_{ij}(D^2u)D_{ij}u_\alpha = 0$$

now in non divergence form. As long as we do not have further information on D^2u, we must think again of F_{ij} as bounded measurable coefficients.

The De Giorgi type theorem for $a_{ij}(x)D_{ij}u_\alpha = 0$ is due to Krylov and Safanov, and states again that solutions of such an equation are Hölder continuous.

We point out that, again this result has "jumped" invariance classes. Rescaling of $a_{ij}(x)$ does not improve them. Unfortunately, this is not enough to "bootstrap", as in the divergence case: The coefficients, $A_{ij}(x) = F_{ij}(D^2u)$, depend on second derivatives. If we will manage to prove that D^2u is Hölder continuous, then, from equation (1), $D_\alpha u$ would be $C^{2,\alpha}(\bar{U})$, i.e., u would be $C^{3,\alpha}(\bar{U})$ and we could improve and improve.

To prove this, once more convexity reappears. If $F(D^2u)$ is concave (or convex) then all pure second derivatives are sub (or super) solutions of the linearized operator. This, together with the fact that D^2u lies in the surface $F(D^2u)$, implies the Hölder continuity of D^2u, and, by the bootstrapping argument u is as smooth as F allows.

The Monge-Ampere equation and optimal transportation

We would like now to turn our attention to the Monge-Ampere equation

$$\det D^2u = \prod \lambda_i = f(x, u, \nabla u) \ .$$

As pointed out before, the equation fits in the context of elliptic equations provided that we consider convex solutions. That is, provided that f is positive. Further $\log\det D^2u = \sum \log\lambda_i$ is concave as function of the λ_i and thus is a concave function of D^2u. Unfortunately $\det D^2u$ *is not* uniformly strictly convex.

For instance if we prescribe

$$\det D^u = \prod \lambda_i = 1$$

ellipticity deteriorates as one of the λ's goes to infinity and some other is forced to go to zero. This difficulty is compensated by two fundamental facts.

1) The rich family of invariances that the Monge-Ampere equation enjoys.
2) Its "hidden" divergence structure.

The divergence structure is due to the fact that $\det D^2u$ can be thought of as the Jacobian of the gradient map: $X \to \nabla u$. Thus for any domain $\bar{U}$

$$\int_{\bar{U}} \det D^2u \, dx = \mathrm{Vol}(\nabla u(\bar{U})).$$

But if $\bar{U} \subset\subset U$, u being convex implies that

$$(\nabla u)|_{\bar{U}} \leq C \operatorname{osc} u|_U.$$

This gives us a sort of "energy inequality" that controls a positive quantity of D^2u by the oscillation of u:

$$\int_{\bar{U}} \det D^2u \leq C(\bar{U}, U)(\operatorname{osc} u)^n.$$

Invariances

The Monge-Ampere equation is invariant of course, under the the standard families of transformations:

a) Rigid motions, R:
$$\det D^2 u(Rx) = f(Rx),$$
b) Translations:
$$\det D^2 u(x+v) = f(x+v),$$
c) Quadratic dialations:
$$\det D^2 \frac{1}{t^2} u(tx) = f(tx).$$

But also

d) Monge-Ampere is invariant under any affine transformation A, of determinant one:
$$\det D^2 u(Ax) = f(Ax)\ .$$

If f is, for instance, in one of the following classes:

a) f constant,
b) f close to constant ($|f-1| \leq \varepsilon$),
c) f bounded away from zero and infinity ($0 < \frac{1}{\sigma} \leq f \leq \sigma$),

any of the transformations above gives a new u in the *same class* of solutions.

For instance, if u is a solution of
$$\det D^2 u = 1$$
then, $u(\varepsilon x, \frac{1}{\varepsilon} y)$ is also a solution of the *same equation.* But this has dramatically "deformed" the graph of u. It is then almost unavoidable that there are singular solutions (Pogorelov).

In fact, for $n \geq 3$, one can construct convex solutions u that contain a line their graph and are not differentiable in the direction transversal to that line, solutions of
$$\det D^2 u = f(x)$$
with f a smooth positive function.

Fortunately, this geometry can only be inherited from the boundary of the domain.

Theorem 0.1. *If in the domain $U \subset R^n$*

a) $\frac{1}{\sigma} \leq \det D^2 u \leq \sigma$,
b) $u \geq 0$,
c) The set $\Gamma = \{u = 0\}$ is not a point, then Γ is generated as "convex combinations" of its boundary points
$$\Gamma = \text{ convex envelope of } \Gamma \cap \partial U\ .$$

A corollary of this theorem is that

a) If we can "cut a slice" of the graph of u, with a hyperplane $l(x)$ so that the support S of $(u-l)^-$ is compactly contained in U, then u is, inside S, both $C^{1,\alpha}$ regular and also $C^{1,\alpha}$- strictly convex, i.e., separates from any of its supporting planes with polynomial growth.

This is the equivalent of De Giorgi's and Krylov-Safanov result (remember that the C^α theorems were applied to the *derivatives* of the solutions of the non-linear equations under consideration).

Note that by an affine transformation and a dilation we can always renormalize the support of the "slice" S to be equivalent to the unit ball of R^n: $B_1 \subset S \subset B_n$.

After this normalization, it is possible to reproduce for u all the classical estimates we had for the Laplacian:

a) (Calderon-Zygmund). If f is close to constant ($|f-1| < \varepsilon$), then $D^2u \in L^p(B_{1/2})$ ($p = p(\varepsilon)$ goes to infinity when ε goes to zero).
b) If $f \in C^{k,\alpha}$ (has up to k derivatives Hölder continuous) then $u \in C^{k+2,\alpha}$ (all second derivatives of u are $C^{k,\alpha}$.

Note that f plays, for Monge-Ampere, simultaneously the role of "right hand side" and "coefficients" due to the structure of its non-linearity.

The Monge-Ampere equation and optimal transportation (the Monge problem)

The Monge-Ampere equation has many applications, not only in geometry, but also in applied areas: optimal design of antenna arrays, vision, statistical mechanics, front formation in meteorology, financial mathematics.

Many of these applications are related to optimal transportation and the Wasserstein metric between probability distributions. In the discrete case, optimal transportation consists of the following.

We are given two sets of k points in R^n: $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$, and want to map the X's onto the Y's, i.e., we look at all one-to-one functions $Y(X_j)$. But we want to do so, minimizing some transportation costs

$$\mathcal{C} = \sum_j C\Big(Y(X_j) - X_j\Big) .$$

For our discussion $C(X-Y) = \frac{1}{2}|X-Y|^2$. It is easy to see that the minimizing map must be the gradient (subdifferential) of a convex potential φ.

In the continuous case, instead of having k-points we have two probability densities, $f(X)\,dX$ and $g(Y)\,dY$ and we want to consider those (admissible) maps $Y(X)$ that "push forward" f to g.

Heuristically that means that in the change of variable formula, we can substitute

$$g(Y(X)) \det D_X Y(X) \text{“=”} f(X).$$

A weak formulation, substitutes the map $Y(X)$, by a joint probability density $\nu(X,Y)$ with marginals $f(X)\,dX$ and $g(Y)\,dY$, i.e.,

$$f(X_0) = \int d_Y \nu(X_0, Y),$$
$$g(Y_0) = \int d_X \nu(X, Y_0).$$

(We don't ask the "map" to be one-to-one any more, the image of X_0 may now spread among "many Y's".

Among all such ν, we want to maximize correlation

$$\mathcal{K} = \int \langle X, Y \rangle d\nu(X,Y)$$

or minimize cost

$$\mathcal{C} = \int \frac{1}{2}|X-Y|^2\, d\nu(X,Y),$$

$\sqrt{\mathcal{C}}$ defines a metric, the Wasserstein metric among probability densities.

Under mild hypothesis, we have the

Theorem 0.2. *The unique optimal ν_0 concentrates in a graph (is actually a one-to-one map, $Y(X)$). Further $Y(X)$ is the subdifferential of a convex potential φ, i.e., $Y(X) = \nabla\varphi$. Heuristically, then, φ must satisfy the Monge-Ampere equation*

$$g(\nabla\varphi)\det D^2\varphi = f(X).$$

For several reasons, the weak theory does not apply in general, but one can still prove, for instance:

Theorem 0.3. *If f and g never vanish or if the supports of f and g are convex sets, the map $Y(X)$ is "one derivative better" than f and g.*

Some applications and current issues

a) It was pointed out by Otto, that the Wasserstein metric can be used to describe the evolution of several of the classical "diffusion" equations: heat equation, porous media, lubrication.

The idea is that a diffusion process for one equation with conservation of mass, consists of the balance of two factors: trying to minimize distance between consecutive distributions ($u(x,t_k)$ and $u(x,t_{k+1})$), plus trying to flatten or smooth (diffuse), $u(x,t_{k+1})$.

This fact has allowed to prove rates of decay to equilibrium in many of the classical equations, as well as a number of new phenomena. The fine relations between the discrete and continuous problems is an evolving issue (rate of convergence, regularity of the discrete problems, etc.).

b) Another family of problems, coming both from geometry and optimal transportation concerns the study of several issues on solutions of Monge-Ampere equations in periodic or random media.

b_1) Liouville type theorems: We start with a theorem of Calabi of Liouville type: Given a global convex solution of Monge-Ampere equation, $\det D^2 u = 1$, u must be a quadratic polynomial. Suppose now that instead of RHS equal to one, we have a general RHS, $f(x)$. Given a global solution, to discover its behavior at infinity we may try to "shrink it" through quadratic transformations:

$$u_\varepsilon = \varepsilon^2 u\left(\frac{x}{\varepsilon}\right) \ , \quad \text{satisfies} \quad \det D^2 u_\varepsilon = \left(\frac{x}{\varepsilon}\right) \ .$$

Suppose now that f averages out at infinity, for instance f is periodic. Then due to the "divergence structure" of Monge-Ampere u_ε should converge to a quadratic polynomial.

Theorem 0.4. *Given a RHS $f(x)$, periodic, with average $\fint f = a$*

i) Given any quadratic polynomial P with $\det D^2 P = a$, there exists a unique periodic function w, such that

$$\det D^2(P + w) = f(x)$$

(w is a "corrector" in homogenization language).

ii) Conversely (Liouville type theorem): Given a global solution u, it must be of the form $P + w$.

What are the implications for homogenization? What can we say if $f(X, u, \nabla u)$ is periodic in X and u? What can we say if $f_\omega(x)$ is random in X?

b_2) Vorticity transport: (2 dimensions) Again in the periodic context we seek a "vorticity density", $\rho(X, t)$ periodic in X. At each time t, ρ generates a periodic "stream function", $\psi(X, t)$ by the equation

$$\det(I + D^2\psi) = \rho \ .$$

In turn, ψ generates a periodic velocity field $v = -(\psi_y, \psi_x)$ that transports ρ:

$$\rho_t + \operatorname{div}(v\rho) = 0 \ .$$

Given some initial data $\rho_0(x)$, what can we say about ρ?

If ρ_0 is a vorticity patch, $\rho_0(x) = 1 + \chi_\Omega$, does it stay that way?

If we choose ρ_0, ψ_0 so that $\rho_0 = F(\psi_0)$, that is $\det I + D^2\psi_0 = F(\psi_0)$, we have a stationary vorticity array, i.e., $\rho(X, t) \equiv \rho_0$.

What can we say, in parallel to the classic theory of rotating fluids, or plasma, where det is substituted by $\Delta\psi$?

c) Another area of research relates to optimal transportation as a natural "map" between probability densities. It has been shown that optimal transportation explains naturally interpolation properties of densities (of Brunn Minkowski type), monotonicity properties (like correlation inequalities that express in which way the probability density, g, is shifted in some cone of directions with respect to f), and concentration properties of g versus f (in which sense for instance, a log concave perturbation of a Gaussian is more concentrated than a Gaussian).

Of particular interest would be to understand optimal transportation as dimension goes to infinity. Since convex potentials are very stable objects, this would provide, under some circumstances, an "infinite dimensional" change of variables formula between probability densities.

d) Finally, one of my favorite problems is to understand the geometry of optimal transportation in the case in which the cost function $C(X-Y)$ is still strictly convex, but not quadratic. In that case, the optimal map is still related to a potential that satisfies

$$\det(I + D(F_j(\nabla\psi))) = \cdots$$

where F_j is now the gradient of the convex conjugate to C.

At this point, we have come full circle and we are now in a higher hierarchy, in a sort of Lagrangian version of the Euler-Lagrange equation from the calculus of variations.

In fact if we put an epsilon in front of D and linearize,

$$\det(I+\varepsilon D(F_j(\nabla\psi))) = 1+\varepsilon \text{ Trace}(D(F_j(\nabla\psi)))+O(\varepsilon^2) = 1+\varepsilon \text{ div } F_j(\nabla\psi) +O(\varepsilon^2).$$

Bibliographical references can be found in the books of J. Gilbarg-N. Trudinger, L.C. Evans and L.A. Caffarelli-X. Cabre for nonlinear PDE's; T. Aubin, I. Bakelman and C. Gutierrez for the Monge-Ampere equation, and the recent surveys by L. Ambrosio and C. Villani for optimal transportation.

ICM 2002 · Vol. I · 189–207

Non-linear Partial Differential Equations in Conformal Geometry*

Sun-Yung Alice Chang† Paul C. Yang‡

0. Introduction

In the study of conformal geometry, the method of elliptic partial differential equations is playing an increasingly significant role. Since the solution of the Yamabe problem, a family of conformally covariant operators (for definition, see section 2) generalizing the conformal Laplacian, and their associated conformal invariants have been introduced. The conformally covariant powers of the Laplacian form a family P_{2k} with $k \in \mathbb{N}$ and $k \leq \frac{n}{2}$ if the dimension n is even. Each P_{2k} has leading order term $(-\Delta)^k$ and is equal to $(-\Delta)^k$ if the metric is flat.

The curvature equations associated with these P_{2k} operators are of interest in themselves since they exhibit a large group of symmetries. The analysis of these equations is of necessity more complicated, it typically requires the derivation of an optimal Sobolev or Moser-Trudinger inequality that always occur at a critical exponent. A common feature is the presence of blowup or bubbling associated to the noncompactness of the conformal group. A number of techniques have been introduced to study the nature of blowup, resulting in a well developed technique to count the topological degree of such equations.

The curvature invariants (called the Q-curvature) associated to such operators are also of higher order. However, some of the invariants are closely related with the Gauss-Bonnet-Chern integrand in even dimensions, hence of intrinsic interest to geometry. For example, in dimension four, the finiteness of the Q-curvature integral can be used to conclude finiteness of topology. In addition, the symmetric functions of the Ricci tensor appear in natural fashion as the lowest order terms of these curvature invariants, these equations offer the possibility to analyze the Ricci tensor itself. In particular, in dimension four the sign of the Q-curvature integral can be used to conclude the sign of the Ricci tensor. Therefore there is ample motivation for the study of such equations.

*Research of Chang is supported in part by NSF Grant DMS-0070542. Research of Yang is supported in part by NSF Grant DMS-0070526.

†Department of Mathematics, Princeton University, Princeton, NJ 08544, USA. E-mail: chang@math.princeton.edu

‡Department of Mathematics, Princeton University, Princeton, NJ 08544, USA. E-mail: yang@math.princeton.edu

In the following sections we will survey some of the development in the area that we have been involved. We gratefully acknowledge the collaborators that we were fortunate to be associated with.

1. Prescribing Gaussian curvature on compact surfaces and the Yamabe problem

In this section we will describe some second order elliptic equations which have played important roles in conformal geometry.

On a compact surface (M, g) with a Riemannian metric g, a natural curvature invariant associated with the Laplace operator $\Delta = \Delta_g$ is the Gaussian curvature $K = K_g$. Under the conformal change of metric $g_w = e^{2w}g$, we have

$$-\Delta w + K = K_w e^{2w} \text{ on } M \tag{1.1}$$

where K_w denotes the Gaussian curvature of (M, g_w). The classical uniformization theorem to classify compact closed surfaces can be viewed as finding solution of equation (1.1) with $K_w \equiv -1$, 0, or 1 according to the sign of $\int K dv_g$. Recall that the Gauss-Bonnet theorem states

$$\int_M K_w \, dv_{g_w} = 2\pi \, \chi(M) \tag{1.2}$$

where $\chi(M)$ is the Euler characteristic of M, a topological invariant. The variational functional with (1.1) as Euler equation for $K_w = \textit{constant}$ is thus given by

$$J[w] = \int_M |\nabla w|^2 dv_g + 2\int_M K w dv_g - (\int_M K dv_g) \log \frac{\int_M dv_{g_w}}{\int_M dv_g}. \tag{1.3}$$

When the surface (M, g) is the standard 2-sphere S^2 with the standard canonical metric, the problem of prescribing Gaussian curvature on S^2 is commonly known as the Nirenberg problem. For general compact surface M, Kazdan and Warner ([57]) gave a necessary and sufficient condition for the function when $\chi(M) = 0$ and some necessary condition for the function when $\chi(M) < 0$. They also pointed out that in the case when $\chi(M) > 0$, i.e. when $(M, g) = (S^2, g_c)$, the standard 2-sphere with the canonical metric $g = g_c$, there is an obstruction for the problem:

$$\int_{S^2} \nabla K_w \cdot \nabla x \; e^{2w} dv_g = 0 \tag{1.4}$$

where x is any of the ambient coordinate function. Moser ([63]) realized that this implicit integrability condition is satisfied if the conformal factor has antipodal symmetry. He proved for an even function f, the only necessary condition for (1.1) to be solvable with $K_w = f$ is that f be positive somewhere. An important tool introduced by Moser is the following inequality ([62]) which is a sharp form of an earlier result of Trudinger ([80]) for the limiting Sobolev embedding of $W_0^{1,2}$ into the Orlicz space e^{L^2}: Let w be a smooth function on the 2-sphere satisfying the

normalizing conditions: $\int_{S^2} |\nabla w|^2 dv_g \leq 1$ and $\bar{w} = 0$ where $\bar{w}$ denotes the mean value of w, then

$$\int_{S^2} e^{\beta w^2} dv_g \leq C \tag{1.5}$$

where $\beta \leq 4\pi$ and C is a fixed constant and 4π is the best constant. If w has antipodal symmetry then the inequality holds for $\beta \leq 8\pi$.

Moser has also established a similar inequality for functions u with compact support on bounded domains in the Euclidean space $\mathbb{R}^n$ with the $W^{1,n}$ energy norm $\int |\nabla u|^n dx$ finite. Subsequently, Carleson and Chang ([14]) found that, contrary to the situation for Sobolev embedding, there is an extremal function realizing the maximum value of the inequality of Moser when the domain is the unit ball in Euclidean space. This fact remains true for simply connected domains in the plane (Flücher [43]), and for some domains in the n-sphere (Soong [77]).

Based on the inequality of Moser and subsequent work of Aubin ([3] and Onofri ([64]), we devised a degree count ([26], [27], [16]) associated to the function f and the Mobius group on the 2-sphere, that is motivated by the Kazdan-Warner condition (1.4). This degree actually computes the Leray-Schauder degree of the equation (1.1) as a nonlinear Fredholm equation. In the special case that f is a Morse function satisfying the condition $\Delta f(x) \neq 0$ at the critical points x of f, this degree can be expressed as:

$$\sum_{\nabla f(q)=0, \Delta f(q)<0} (-1)^{ind(q)} - 1. \tag{1.6}$$

The latter degree count is also obtained later by Chang-Liu ([15]) and Han ([54]).

There is another interesting geometric interpretation of the functional J given by Ray-Singer ([73]) and Polyakov ([71]); (see also Okikiolu [67])

$$J[w] = 12\pi \ \log \left(\frac{\det \ \Delta_g}{\det \ \Delta_{g_w}}\right) \tag{1.7}$$

for metrics g_w with the volume of g_w equals the volume of g; where the determinant of the Laplacian $det\ \Delta_g$ is defined by Ray-Singer via the "regularized" zeta function. In [64], (see also Hong [55]), Onofri established the sharp inequality that on the 2-sphere $J[w] \geq 0$ and $J[w] = 0$ precisely for conformal factors w of the form $e^{2w} g_0 = T^* g_0$ where T is a Mobius transformation of the 2-sphere. Later Osgood-Phillips-Sarnak ([65], [66]) arrived at the same sharp inequality in their study of heights of the Laplacian. This inequality also plays an important role in their proof of the C^∞ compactness of isospectral metrics on compact surfaces.

The formula of Polyakov-Ray-Singer has been generalized to manifolds of dimension greater than two in many different settings; one of which we will discuss in section 2 below. There is also a general study of extremal metrics for $det\ \Delta_g$ or $det\ L_g$ for metrics g in the same conformal class with a fixed volume or for all metrics with a fixed volume([5], [8], [7], [72], [68]). A special case of the remarkable results of Okikiolu ([68]) is that among all metrics with the same volume as the standard metric on the 3-sphere, the standard canonical metric is a local maximum for the functional $det\ \Delta_g$.

More recently, there is an extensive study of a generalization of the equation (1.1) to compact Riemann surfaces. Since Moser's argument is readily applicable to a compact surface (M, g), a lower bound for similarly defined functional J on (M, g) continues to hold in that situation. The Chern-Simons-Higgs equation in the Abelian case is given by:

$$\Delta w = \rho e^{2w}(e^{2w} - 1) + 2\pi \sum_{i=1}^{N} \delta_{p_i}. \tag{1.8}$$

A closely related equation is the mean field equation:

$$\Delta w + \rho(\frac{he^{2w}}{\int he^{2w}} - 1) = 0, \tag{1.9}$$

where ρ is a real parameter that is allowed to vary.

There is active development on these equations by several group of researchers including ([13], [36], [79], [78], [31]).

On manifolds (M^n, g) for n greater than two, the conformal Laplacian L_g is defined as $L_g = -c_n\Delta_g + R_g$ where $c_n = \frac{4(n-1)}{n-2}$, and R_g denotes the scalar curvature of the metric g. An analogue of equation (1.1) is the equation, commonly referred to as the Yamabe equation, which relates the scalar curvature under conformal change of metric to the background metric. In this case, it is convenient to denote the conformal metric as $\bar{g} = u^{\frac{4}{n-2}} g$ for some positive function u, then the equation becomes

$$L_g u = \bar{R}\, u^{\frac{n+2}{n-2}}. \tag{1.10}$$

The famous Yamabe problem to solve (1.10) with $\bar{R}$ a constant has been settled by Yamabe ([85]), Trudinger ([81]), Aubin ([2]) and Schoen ([74]). The corresponding problem to prescribe scalar curvature has been intensively studied in the past decades by different groups of mathematicians, we will not be able to survey all the results here. We will just mention that the degree theory for existence of solutions on the n-sphere has been achieved by Bahri-Coron ([4]), Chang-Gursky-Yang ([16]) and Schoen-Zhang ([75]) for $n = 3$ and under further constraints on the functions for $n \geq 4$ by Y. Li ([59]) and by C-C. Chen and C.-S. Lin ([32]).

2. Conformally covariant differential operators and the Q-curvatures

It is well known that in dimension two, under the conformal change of metrics $g_w = e^{2w} g$, the associated Laplacians are related by

$$\Delta_{g_w} = e^{-2w}\Delta_g. \tag{2.1}$$

Similarly on (M^n, g), the conformal Laplacian $L = -\frac{4(n-1)}{n-2}\Delta + R$ transforms under the conformal change of metric $\bar{g} = u^{\frac{4}{n-2}} g$:

$$L_{\bar{g}} = u^{-\frac{n+2}{n-2}} L_g(u\cdot) \tag{2.2}$$

In general, we call a metrically defined operator A conformally covariant of bidegree (a, b), if under the conformal change of metric $g_\omega = e^{2\omega} g$, the pair of corresponding operators A_ω and A are related by

$$A_\omega(\varphi) = e^{-b\omega} A(e^{a\omega}\varphi) \quad \text{for all} \quad \varphi \in C^\infty(M^n)\,.$$

Note that in this notation, the conformal Laplacian opertor is conformally covariant of bidegree $(\frac{n-2}{2}, \frac{n+2}{2})$.

There are many operators besides the Laplacian Δ on compact surfaces and the conformal Laplacian L on general compact manifold of dimension greater than two which have the conformal covariance property. We begin with the fourth order operator on 4-manifolds discovered by Paneitz ([70]) in 1983 (see also [37]):

$$P\varphi \equiv \Delta^2\varphi + \delta\left(\frac{2}{3}Rg - 2\mathrm{Ric}\right) d\varphi$$

where δ denotes the divergence, d the deRham differential and Ric the Ricci tensor of the metric. The *Paneitz* operator P (which we will later denote by P_4) is conformally covariant of bidegree $(0, 4)$ on 4-manifolds, i.e.

$$P_{g_w}(\varphi) = e^{-4\omega} P_g(\varphi) \quad \text{for all} \quad \varphi \in C^\infty(M^4)\,.$$

More generally, T. Branson ([6]) has extended the definition of the fourth order operator to general dimensions $n \neq 2$; which we call the conformal Paneitz operator:

$$P_4^n = \Delta^2 + \delta\left(a_n Rg + b_n \mathrm{Ric}\right) d + \frac{n-4}{2} Q_4^n \tag{2.3}$$

where

$$Q_4^n = c_n |Ric|^2 + d_n R^2 - \frac{1}{2(n-1)} \Delta R, \tag{2.4}$$

and

$$a_n = \frac{(n-2)^2 + 4}{2(n-1)(n-2)},\ b_n = -\frac{4}{n-2},\ c_n = -\frac{2}{(n-2)^2},\ d_n = \frac{n^3 - 4n^2 + 16n - 16}{8(n-1)^2(n-2)^2}.$$

The conformal Paneitz operator is conformally covariant of bidegree $(\frac{n-4}{2}, \frac{n+4}{2})$. As in the case of the second order conformally covariant operators, the fourth order Paneitz operators have associated fourth order curvature invariants Q: in dimension $n = 4$ we write the conformal metric $g_w = e^{2w} g$; $Q = Q_g = \frac{1}{2}(Q_4^4)_g$, then

$$Pw + 2Q \;=\; 2Q_{g_w} e^{4w} \tag{2.5}$$

and in dimensions $n \neq 1, 2, 4$ we write the conformal metric as $\bar{g} = u^{\frac{4}{n-4}} g$:

$$P_4^n u \;=\; \bar{Q}_4^n u^{\frac{n+4}{n-4}}. \tag{2.6}$$

In dimension $n = 4$ the Q-curvature equation is closely connected to the Gauss-Bonnet-Chern formula:

$$4\pi^2 \chi(M^4) = \int (Q + \frac{1}{8}|W|^2)\, dv \tag{2.7}$$

where W denotes the Weyl tensor, and the quantity $|W|^2 dv$ is a pointwise conformal invariant. Therefore the Q-curvature integral $\int Q dv$ is a conformal invariant. The basic existence theory for the Q-curvature equation is outlined in [28]:

Theorem 2.1. *If $\int Q dv < 8\pi^2$ and the P operator is positive except for constants, then equation (2.5) may be solved with Q_{g_w} given by a constant.*

It is remarkable that the conditions in this existence theorem are shown by M. Gursky ([51]) to be a consequence of the assumptions that (M, g) has positive Yamabe invariant[1], and that $\int Q dv > 0$. In fact, he proves that under these conditions P is a positive operator and $\int Q dv \leq 8\pi^2$ and that equality can hold only if (M, g) is conformally equivalent to the standard 4-sphere. This latter fact may be viewed as the analogue of the positive mass theorem that is the source for the basic compactness result for the Q-curvature equation as well as the associated fully nonlinear second order equations that we discuss in section 4. Gursky's argument is based on a more general existence result in which we consider a family of 4-th order equations

$$\gamma_1 |W|^2 + \gamma_2 Q - \gamma_3 \Delta R = \bar{k} \cdot \mathrm{Vol}^{-1} \tag{2.8}$$

where $\bar{k} = \int (\gamma_1 |W|^2 + \gamma_2 Q) dv$. These equations typically arise as the Euler equation of the functional determinants. For a conformally covariant operator A of bidegree (a, b) with $b - a = 2$ Branson and Orsted ([9]) gave an explicit computation of the normalized form of $\log \frac{\det A_w}{\det A}$ which may be expressed as:

$$F[w] = \gamma_1 I[w] + \gamma_2 II[w] + \gamma_3 III[w] \tag{2.9}$$

where $\gamma_1, \gamma_2, \gamma_3$ are constants depending only on A and

$$I[w] = 4 \int |W|^2 w dv - \left(\int |W|^2 dv \right) \log \frac{\int e^{4w} dv}{\int dv},$$

$$II[w] = \langle Pw, w \rangle + 4 \int Q w dv - \left(\int Q dv \right) \log \frac{\int e^{4w} dv}{\int dv},$$

$$III[w] = \frac{1}{3} \left(\int R_{g_w}^2 dv_{g_w} - \int R^2 dv \right).$$

In [28], we gave the general existence result:

Theorem 2.2. *If the functional F satisfies $\gamma_2 > 0$, $\gamma_3 > 0$, and $\bar{k} < 8\gamma_2 \pi^2$, then $\inf_{w \in W^{2,2}} F[w]$ is attained by some function w_d and the metric $g_d = e^{2w_d} g_0$ satisfies the equation*

$$\gamma_1 |W|^2 + \gamma_2 Q_d - \gamma_3 \triangle_d R_d = \bar{k} \cdot Vol(g_d)^{-1}. \tag{2.10}$$

Furthermore, g_d is smooth.

[1] The Yamabe invariant $Y(M, g)$ is defined to be $Y(M, g) \equiv \inf_w \frac{\int_M R_{g_w} dv_{g_w}}{vol(g_w)^{\frac{n-2}{n}}}$; where n denotes the dimension of M. $Y(M, g)$ is confomally invariant and the sign of $Y(M, g)$ agrees with that of the first eigenvalue of L_g.

This existence result is based on extensions of Moser's inequality by Adams ([1], on manifolds [40]) to operators of higher order. In the special case of (M^4, g), the inequality states that for functions in the Sobolev space $W^{2,2}(M)$ with $\int_M (\Delta w)^2 dv_g \leq 1$, and $\bar{w} = 0$, we have

$$\int_M e^{32\pi^2 w^2} dv_g \leq C, \tag{2.11}$$

for some constant C. The regularity for minimizing solutions was first given in [17], and later extended to all solutions by Uhlenbeck and Viaclovsky ([82]). There are several applications of these existence result to the study of conformal structures in dimension $n = 4$. In section 4 we will discuss the use of such fourth order equation as regularization of the more natural fully nonlinear equation concerned with the Weyl-Schouten tensor. Here we will mention some elegant application by M. Gursky ([50]) to characterize a number of extremal conformal structures.

Theorem 2.3. *Suppose (M, g) is a compact oriented manifold of dimension four with positive Yamabe invariant.*

(i) If $\int Q_g dv_g = 0$, and if M admits a non-zero harmonic 1-form, then (M, g) is conformal equivalent to a quotient of the product space $S^3 \times \mathbb{R}$. In particular (M, g) is locally conformally flat.

(ii) If $b_2^+ > 0$ (i.e. the intersection form has a positive element), then with respect to the decompostion of the Weyl tensor into the self dual and anti-self dual components $W = W^+ \oplus W^-$,

$$\int_M |W_g^+|^2 dv_g \geq \frac{4\pi^2}{3}(2\chi + 3\tau), \tag{2.12}$$

where τ is the signature of M. Moreover the equality holds if and only if g is conformal to a (positive) Kahler-Einstein metric.

In dimensions higher than four, the analogue of the Yamabe equation for the fourth order Paneitz equation is being investigated by a number of authors. In particular, Djadli-Hebey-Ledoux ([34]) studied the question of coercivity of the operators P as well as the positivity of the solution functions, Djadli-Malchiodi-Ahmedou ([35]) have studied the blowup analysis of the Paneitz equation. In dimension three, the fourth order Paneitz equation involves a negative exponent, there is now an existence result ([84]) in case the Paneitz operator is positive.

In general dimensions there is an extensive theory of local conformal invariants according to the theory of Fefferman and Graham ([41]). For manifolds of general dimension n, when n is even, the existence of a n-th order operator P_n conformally covariant of bidegree $(0, n)$ was verified in [45]. However it is only explicitly known on the standard Euclidean space $\mathbb{R}^n$ and hence on the standard sphere S^n. For all n, on (S^n, g), there also exists an n-th order (pseudo) differential operator $\mathbb{P}_n$ which is the pull back via sterographic projection of the operator $(-\Delta)^{n/2}$ from $\mathbb{R}^n$ with Euclidean metric to (S^n, g). $\mathbb{P}_n$ is conformally covariant of bi-degree (0, n), i.e. $(\mathbb{P}_n)_w = e^{-nw}\mathbb{P}_n$. The explicit formulas for $\mathbb{P}_n$ on S^n has been computed in

Branson ([7]) and Beckner ([5]):

$$\begin{cases} \text{For } n \text{ even} & \mathbb{P}_n = \prod_{k=0}^{\frac{n-2}{2}}(-\Delta + k(n-k-1)), \\ \text{For } n \text{ odd} & \mathbb{P}_n = \left(-\Delta + \left(\frac{n-1}{2}\right)^2\right)^{1/2} \prod_{k=0}^{\frac{n-3}{2}}(-\Delta + k(n-k-1)). \end{cases} \tag{2.13}$$

Using the method of moving planes, it is shown in [29] that all solutions of the (pseudo-) differential equation:

$$\mathbb{P}_n w + (n-1)! = (n-1)! e^{nw} \tag{2.14}$$

are given by actions of the conformal group of S^n. As a consequence, we derive ([28]) the sharp version of a Moser-Trudinger inequality for spheres in general dimensions. This inequality is equivalent to Beckner's inequality ([5]).

$$\log \frac{1}{|S^n|}\int_{S^n} e^{nw} dv \leq \frac{1}{|S^n|}\int_{S^n} (nw + \frac{n}{2(n-1)!} w\mathbb{P}_n(w)) dv, \tag{2.15}$$

and equality holds if and only if e^{nw} represents the Jacobian of a conformal transformation of S^n.

In a recent preprint, S. Brendle is able to derive a general existence result for the prescribed Q-curvature equation under natural conditions:

Theorem 2.4. ([10]) *For a compact manifold (M^{2m}, g) satisfying*

(i) P_{2m} be positive except on constants,

(ii) $\int_M Q_g dv_g < C_{2m}$ where C_{2m} represents the value of the corresponding Q-curvature integral on the standard sphere (S^{2m}, g_c), the equation $P_{2m} w + Q = Q_w e^{2mw}$ has a solution with Q_w given by a constant.

Brendle's remarkable argument uses a $2m$-th order heat flow method in which again inequality of Adams ([1]) (the only available tool) is used.

In another recent development, the n-th order Q-curvature integral can be interpreted as a renormalized volume of the conformally compact manifold (N^{n+1}, h) of which (M^n, g) is the conformal infinity. In particular, Graham-Zworski ([46]) and Fefferman-Graham ([42]) have given in the case n is an even integer, a spectral theory interpretation to the n-th order Q-curvature integral that is intrinsic to the boundary conformal structure. In the case n is odd, such an interpretation is still available, however it may depend on the conformal compactification.

3. Boundary operator, Cohn-Vossen inequality

To develop the analysis of the Q-curvature equation, it is helpful to consider the associated boundary value problems. In the case of compact surface with boundary (N^2, M^1, g) where the metric g is defined on $N^2 \cup M^1$; the Gauss-Bonnet formula becomes

$$2\pi\chi(N) = \int_N K\, dv + \oint_M k\, d\sigma, \tag{3.1}$$

where k is the geodesic curvature on M. Under conformal change of metric g_w on N, the geodesic curvature changes according to the equation

$$\frac{\partial}{\partial n}w + k \;=\; k_w e^w \;\text{ on M.} \tag{3.2}$$

Ray-Singer-Polyakov log-determinant formula has been generalized to compact surface with boundary and the extremal metric of the formula has been studied by Osgood-Phillips-Sarnak ([66]). The role played by the Onofri inequality is the classical Milin-Lebedev inequality:

$$\log \oint_{S^1} e^{(w-\bar{w})}\frac{d\theta}{2\pi} \leq \frac{1}{4}\left(\int_D w(-\Delta w)\frac{dx}{\pi} \;+\; 2\oint_{S^1} w\frac{\partial w}{\partial n}\frac{d\theta}{2\pi}\right), \tag{3.3}$$

where D is the unit disc on $\mathbb{R}^2$ with the flat metric dx, and n is the unit outward normal.

One can generalize above results to four manifold with boundary (N^4, M^3, g); with the role played by $(-\Delta, \frac{\partial}{\partial n})$ replaced by (P_4, P_3) and with (K, k) replaced by (Q, T); where P_4 is the Paneitz opertor and Q the curvature discussed in section 2; and where P_3 is the boundary operator constructed by Chang-Qing ([22]). The key property of P_3 is that it is conformally covariant of bidegree $(0, 3)$, when operating on functions defined on the boundary of compact 4-manifolds; and under conformal change of metric $\bar{g} = e^{2w}g$ on N^4 we have at the boundary M^3

$$P_3 w + T \;=\; T_w e^{3w}. \tag{3.4}$$

We refer the reader to [22] for the precise definitions of P_3 and T and will here only mention that on (B^4, S^3, dx), where B^4 is the unit ball in $\mathbb{R}^4$, we have

$$P_4 = (-\Delta)^2,\; P_3 = -\left(\frac{1}{2}\,\frac{\partial}{\partial n}\,\Delta + \tilde{\Delta}\frac{\partial}{\partial n} + \tilde{\Delta}\right) \quad \text{and} \;\; T = 2, \tag{3.5}$$

where $\tilde{\Delta}$ is the intrinsic boundary Laplacian on M.

In this case the Gauss-Bonnet-Chern formula may be expressed as:

$$4\pi^2\chi(N) = \int_N (\,Q \,+ \frac{1}{8}|W|^2)\; dv \;+ \oint_M (T + \mathcal{L})\; d\sigma, \tag{3.6}$$

where $\mathcal{L}$ is a third order boundary curvature invariant that t ransforms by scaling under conformal change of metric. The analogue of the sharp form of the Moser-Trudinger inequality for the pair $(B^4, S^3,\, dx)$ is given by the following analogue of the Milin-Lebedev inequality:

Theorem 3.1. ([23]) *Suppose* $w \in C^\infty(\bar{B}^4)$. *Then*

$$\begin{aligned} &\log\left\{\frac{1}{2\pi^2}\oint_{S^3} e^{3(w-\bar{w})}d\sigma\right\} \\ \leq\;& \frac{3}{16\pi^2}\left\{\int_{B^4} w\Delta^2 w dx \;+ \oint_{S^3}\left(2wP_3 w - \frac{\partial w}{\partial n} + \frac{\partial^2 w}{\partial n^2}\right)\, d\sigma\right\}, \end{aligned} \tag{3.7}$$

under the boundary assumptions $\frac{\partial w}{\partial n}|_{S^3} = e^w - 1$ *and* $\int_{S^3} R_w d\sigma_{g_w} = \int_{S^3} R d\sigma$ *where* R *is the scalar curvature of* S^3. *Moreover the equality holds if and only if* $e^{2w}dx$ *on* B^4 *is isometric to the standard metric via a conformal transformation of the pair* (B^4, S^3, dx).

The boundary version (3.6) of the Gauss-Bonnet-Chern formula can be used to give an extension of the well known Cohn-Vossen-Huber formula. Let us recall ([33], [56]) that a complete surface (N^2, g) with Gauss curvature in L^1 has a conformal compactification $\bar{N} = N \cup \{q_1, ..., q_l\}$ as a compact Riemann surface and

$$2\pi\chi(N) = \int_N K dA + \sum_{k=1}^{l} \nu_k, \tag{3.8}$$

where at each end q_k, take a conformal coordinate disk $\{|z| < r_0\}$ with q_k at its center, then ν_k represents the following limiting isoperimetric constant:

$$\nu_k = \lim_{r \to 0} \frac{Length(\{|z| = r\})^2}{2Area(\{r < |z| < r_0\})}. \tag{3.9}$$

This result can be generalized to dimension $n = 4$ for locally conformally flat metrics. In general dimensions, Schoen-Yau ([76]) proved that locally conformally flat metrics in the non-negative Yamabe class has injective development map into the standard spheres as domains whose complement have small Hausdorff dimension (at most $\frac{n-2}{2}$). It is possible to further constraint the topology as well as the end structure of such manifolds by imposing the natural condition that the Q-curvature be in L^1.

Theorem 3.2. ([24], [25]) *Suppose* (M^4, g) *is a complete conformally flat manifold, satisfying the conditions:*
(i) The scalar curvature R_g *is bounded between two positive constants and* $\nabla_g R_g$ *is also bounded;*
(ii) The Ricci curvature is bounded below;
(iii) $\int_M |Q_g| dv_g < \infty$;
then
(a) if M *is simply connected, it is conformally equivalent to* $S^4 - \{q_1, ..., q_l\}$ *and we have*

$$4\pi^2 \ \chi(M) = \int_M Q_g \ dv_g \ + \ 4\pi^2 l \ ; \tag{3.10}$$

(b) if M *is not simply connected, and we assume in addition that its fundamental group is realized as a geometrically finite Kleinian group, then we conclude that* M *has a conformal compactification* $\bar{M} = M \cup \{q_1, ..., q_l\}$ *and equation (3.10) holds.*

This result gives a geometric interpretation to the Q-curvature integral as measuring an isoperimetric constant. There are two elements in this argument. The first is to view the Q-curvature integral over sub-level sets of the conformal factors as the second derivative with respect to w of the corresponding volume

integral. This comparison is made possible by making use of the formula (3.4). A second element is an estimate for conformal metrics $e^{2w}|dx|^2$ defined over domains $\Omega \subset \mathbb{R}^4$ satisfying the conditions of Theorem 3.2 must have a uniform blowup rate near the boundary:

$$e^{w(x)} \cong \frac{1}{d(x, \partial\Omega)}. \tag{3.11}$$

This result has an appropriate generalization to higher even dimensional situation, in which one has to impose additional curvature bounds to control the lower order terms in the integral. One such an extension is obtained in the thesis of H. Fang ([39]).

It remains an interesting question how to extend this analysis to include the case when the dimension is an odd integer.

4. Fully nonlinear equations in conformal geometry in dimension four

In dimensions greater than two, the natural curvature invariants in conformal geometry are the Weyl tensor W, and the Weyl-Schouten tensor $A = Ric - \frac{R}{2(n-1)}g$ that occur in the decomposition of the curvature tensor; where Ric denotes the Ricci curvature tensor:

$$Rm = W \oplus \frac{1}{n-2} A \owedge g. \tag{4.1}$$

Since the Weyl tensor W transforms by scaling under conformal change $g_w = e^{2w}g$, only the Weyl-Schouten tensor depends on the derivatives of the conformal factor. It is thus natural to consider $\sigma_k(A_g)$ the k-th symmetric function of the eigenvalues of the Weyl-Schouten tensor A_g as curvature invariants of the conformal metrics. As a differential invariant of the conformal factor w, $\sigma_k(A_{g_w})$ is a fully nonlinear expression involving the Hessian and the gradient of the conformal factor w. We have abbreviating A_w for A_{g_w}:

$$A_w = (n-2)\{-\nabla^2 w + dw \otimes dw - \frac{|\nabla w|^2}{2}\} + A_g. \tag{4.2}$$

The equation

$$\sigma_k(A_w) = 1 \tag{4.3}$$

is a fully nonlinear version of the Yamabe equation. For example, when $k = 1$, $\sigma_1(A_g) = \frac{n-2}{2(n-1)}R_g$, where R_g is the scalar curvature of (M, g) and equation (4.3) is the Yamabe equation which we have discussed in section 1. When $k = 2$, $\sigma_2(A_g) = \frac{1}{2}(|Trace\ A_g|^2 - |A_g|^2) = \frac{n}{8(n-1)}R^2 - \frac{1}{2}|Ric|^2$. In the case when $k = n$, $\sigma_n(A_g) =$ *determinant of* A_g, an equation of Monge-Ampere type. To illustrate that (4.3) is a fully non-linear elliptic equation, we have for example when $n = 4$,

$$\begin{aligned} \sigma_2(A_{g_w})e^{4w} &= \sigma_2(A_g) + 2((\Delta w)^2 - |\nabla^2 w|^2 \\ &\quad + (\nabla w, \nabla|\nabla w|^2) + \Delta w|\nabla w|^2) \\ &\quad + \text{lower order terms}, \end{aligned} \tag{4.4}$$

where all derivative are taken with respect to the g metric.

For a symmetric $n \times n$ matrix M, we say $M \in \Gamma_k^+$ in the sense of Garding ([44]) if $\sigma_k(M) > 0$ and M may be joined to the identity matrix by a path consisting entirely of matrices M_t such that $\sigma_k(M_t) > 0$. There is a rich literature conerning the equation

$$\sigma_k(\nabla^2 u) = f, \tag{4.5}$$

for a positive function f. In the case when $M = (\nabla^2 u)$ for convex functions u defined on the Euclidean domains, regularity theory for equations of $\sigma_k(M)$ has been well established for $M \in \Gamma_k^+$ for Dirichlet boundary value problems by Caffarelli-Nirenberg-Spruck ([12]); for a more general class of fully non-linear elliptic equations not necessarily of divergence form by Krylov ([58]), Evans ([38]) and for Monge-Ampere equations by Pogorelov ([69]) and by Caffarelli ([11]). The Monge-Ampere equation for prescribing the Gauss-Kronecker curvature for convex hypersurfaces has been studied by Guan-Spruck ([47]). Some of the techniques in these work can be modified to study equation (4.3) on manifolds. However there are features of the equation (4.3) that are distinct from the equation (4.5). For example, the conformal invariance of the equation (4.3) introduces a non-compactness due to the action of the conformal group that is absent for the equation (4.5).

When $k \neq \frac{n}{2}$ and the manifold (M, g) is locally conformally flat, Viaclovsky ([83]) showed that the equation (4.3) is the Euler equation of the variational functional $\int \sigma_k(A_{g_w}) dv_{g_w}$. In the exceptional case $k = n/2$, the integral $\int \sigma_k(A_g) dv_g$ is a conformal invariant. We say $g \in \Gamma_k^+$ if the corresponding Weyl-Schouten tensor $A_g(x) \in \Gamma_k^+$ for every point $x \in M$. For $k = 1$ the Yamabe equation (1.10) for prescribing scalar curvature is a semilinear one; hence the condition for $g \in \Gamma_1^+$ is the same as requiring the operator $L_g = -\frac{4(n-1)}{n-2}\Delta_g + R_g$ be a positive operator. The existence of a metric with $g \in \Gamma_k^+$ implies a sign for the curvature functions ([52], [18], [48]).

Proposition 4.1. *On (M^n, g),*
(i) When $n = 3$ and $\sigma_2(A_g) > 0$, then either $R_g > 0$ and the sectional curvature of g is positive or $R_g < 0$ and the sectional curvature of g is negative on M.
(ii) When $n = 4$ and $\sigma_2(A_g) > 0$, then either $R_g > 0$ and $Ric_g > 0$ on M or $R_g < 0$ and $Ric_g < 0$ on M.
(iii) For general n and $A_g \in \Gamma_k^+$ for some $k \geq \frac{n}{2}$, then $Ric_g > 0$.

In dimension 3, one can capture all metrics with constant sectional curvature (i.e. space forms) through the study of σ_2.

Theorem 4.2. ([52]) *On a compact 3-manifold, for any Riemannian metric g, denote $\mathcal{F}_2[g] = \int_M \sigma_2(A_g) dv_g$. Then a metric g with $\mathcal{F}_2[g] \geq 0$ is critical for the functional $\mathcal{F}_2$ restricted to class of metrics with volume one if and only if g has constant sectional curvature.*

The criteria for existence of a conformal metric $g \in \Gamma_k^+$ is not as easy for $k > 1$ since the equation is a fully nonlinear one. However when $n = 4, k = 2$ the invariance of the integral $\int \sigma_2(A_g) dv_g$ is a reflection of the Chern-Gauss-Bonnet

formula

$$8\pi^2\chi(M) = \int_M (\sigma_2(A_g) + \frac{1}{4}|W_g|^2)dv_g. \tag{4.6}$$

In this case it is possible to find a criteria:

Theorem 4.3. ([18])*For a closed 4-manifold (M, g) satisfying the following conformally invariant conditions:*
(i) $Y(M, g) > 0$, and
(ii) $\int \sigma_2(A_g)dv_g > 0$;
then there exists a conformal metric $g_w \in \Gamma_2^+$.

Remark. In dimension four, the condition $g \in \Gamma_2^+$ implies that $R > 0$ and Ricci is positive everywhere. Thus such manifolds have finite fundamental group. In addition, the Chern-Gauss-Bonnet formula and the signature formula shows that this class of 4-manifolds satisfy the same conditions as that of an Einstein manifold with positive scalar curvatures. Thus it is the natural class of 4-manifolds in which to seek an Einstein metric.

The existence result depends on the solution of a family of fourth order equations involving the Paneitz operator ([70]), which we have discussed in section 2. In the following we briefly outline this connection. Recall that in dimension four, the Paneitz operator P a fourth order curvature called the Q-curvature:

$$P_g w + 2Q_g = 2Q_{g_w}e^{4w}. \tag{4.7}$$

The relation between Q and $\sigma_2(A)$ in dimension 4 is given by

$$Q_g = \frac{-1}{12}\Delta R_g + \frac{1}{2}\sigma_2(A_g). \tag{4.8}$$

In view of the existence results of Theorem 2.1 and Theorem 2.2, it is natural to find a solution of

$$\sigma_2(A_g) = f \tag{4.9}$$

for some positive function f. It turns out that it is natural to choose $f = c|W_g|^2$ for some constant c and to use the continuity method to solve the family of equations

$$(*)_\delta : \qquad \sigma_2(A_g) = \frac{\delta}{4}\Delta_g R_g - 2\gamma|W_g|^2 \tag{4.10}$$

where γ is chosen so that $\int \sigma_2(A_g)dv_g = -2\gamma \int |W_g|^2 dv_g$, for $\delta \in (0, 1]$ and let δ tend to zero.

Indeed when $\delta = 1$, solution of (4.10) is a special case of an extremal metric of the log-determinant type functional $F[w]$ in Theorem 2.2, where we choose $\gamma_2 = 1$, $\gamma_3 = \frac{1}{24}$, we then choose $\gamma = \gamma_1$ so that $\bar{k} = 0$. Notice that in this case, the assumption (ii) in the statement of Theorem 4.3 implies that $\gamma < 0$. When $\delta = \frac{2}{3}$, equation (4.10) amounts to solving the equation

$$Q_g = -\gamma|W_g|^2, \tag{4.11}$$

which we can solve by applying Theorem 2.1. Thus the bulk of the analysis consist in obtaining apriori estimates of the solution as δ tends to zero, showing essentially

that in the equation the term $\frac{\delta}{4}\Delta R$ is small in the weak sense. The proof ends by first modifying the function $|W|^2$ to make it strictly positive and by then applying the Yamabe flow to the metrics g_δ to show that for sufficiently small δ the smoothing provided by the Yamabe flow yields a metric $g \in \Gamma_2^+$.

The equation (4.3) becomes meaningful for 4-manifolds which admits a metric $g \in \Gamma_2^+$. In the article ([19]), when the manifold (M, g) is not conformally equivalent to (S^4, g_c), we provide apriori estimates for solutions of the equation (4.9) where f is a given positive smooth function. Then we apply the degree theory for fully non-linear elliptic equation to the following 1-parameter family of equations

$$\sigma_2(A_{g_t}) = tf + (1-t) \tag{4.12}$$

to deform the original metric to one with constant $\sigma_2(A_g)$.

In terms of geometric application, this circle of ideas may be applied to characterize a number of interesting conformal classes in terms of the the relative size of the conformal invariant $\int \sigma_2(A_g)dV_g$ compared with the Euler number.

Theorem 4.4. ([21]) *Suppose (M, g) is a closed 4-manifold with $Y(M, g) > 0$.*

(I) If $\int_M \sigma_2(A_g)dv_g > \frac{1}{4}\int_M |W_g|^2\, dv_g$, then M is diffeomorphic to (S^4, g_c) or $(\mathbb{R}P^4, g_c)$.

(II) If M is not diffeomorphic to (S^4, g_c) or $(\mathbb{R}P^4, g_c)$ and $\int_M \sigma_2(A_g)dv_g = \frac{1}{4}\int_M |W_g|^2\, dv_g$, then either

(a) (M, g) is conformally equivalent to $(\mathbb{C}P^2, g_{FS})$, or

(b) (M, g) is conformal equivalent to $((S^3 \times S^1)/\Gamma, g_{prod})$.

Remark. The theorem above is an L^2 version of an earlier result of Margerin [61]. The first part of the theorem should be compared to a result of Hamilton ([53]); where he pioneered the method of Ricci flow and established the diffeomorphism of M^4 to the 4-sphere under the assumption that the curvature operator be positive.

This first part of Theorem 4.4 applies the existence argument to find a conformal metric g' which satisfies the pointwise inequality

$$\sigma_2(A_{g'}) > \frac{1}{4}|W_{g'}|^2. \tag{4.13}$$

The diffeomorphism assertion follows from Margerin's ([61]) precise convergence result for the Ricci flow: such a metric will evolve under the Ricci flow to one with constant curvature. Therefore such a manifold is diffeomorphic to a quotient of the standard 4-sphere.

For the second part of the assertion, we argue that if such a manifold is not diffeomorphic to the 4-sphere, then the conformal structure realizes the minimum of the quantity $\int |W_g|^2 dv_g$, and hence its Bach tensor vanishes. There are two possibilities depending on whether the Euler number is zero or not. In the first case, an earlier result of Gursky ([50]) shows the metric is conformal to that of the space $S^1 \times S^3$. In the second case, we solve the equation

$$\sigma_2(A_{g'}) = \frac{1-\epsilon}{4}|W_{g'}|^2 + C_\epsilon, \tag{4.14}$$

where C_ϵ is a constant which tend to zero as ϵ tend to zero. We then let ϵ tends to zero. We obtain in the limit a $C^{1,1}$ metric which satisfies the equation on the open set $\Omega = \{x|W(x) \neq 0\}$:

$$\sigma_2(A_{g'}) = \frac{1}{4}|W_{g'}|^2. \tag{4.15}$$

Then a Lagrange multiplier computation shows that the curvature tensor of the limit metric agrees with that of the Fubini-Study metric on the open set where $W \neq 0$. Therefore $|W_{g'}|$ is a constant on Ω thus W cannot vanish at all. It follows from the Cartan-Kahler theory that the limit metric agrees with the Fubini-Study metric of $\mathbb{C}P^2$ everywhere.

There is a very recent work of A. Li and Y. Li ([60]) extending work of ([20]) to classify the entire solutions of the equation $\sigma_k(A_g) = 1$ on $\mathbb{R}^n$ thus providing apriori estimates for this equation in the locally conformally flat case. There is also a very recent work ([49]) on the heat flow of this equation, we have ([30]) used this flow to derive the sharp version of the Moser-Onofri inequality for the $\sigma_{\frac{n}{2}}$ energy for all even dimensional spheres. In general, the geometric implications of the study of σ_k for manifolds of dimension greater than four remains open.

References

[1] D. Adams; *A sharp inequality of J. Moser for higher order derivatives*, Ann. of Math., 128 (1988), 385–398.

[2] T. Aubin; *Equations differentielles non lineaires et probleme de Yamabe concernant la courbure scalaire*, J. Math. Pures Appl. 55 (1976), 269–296.

[3] T. Aubin; *Meilleures constantes dans le theorem d'inclusion de Sobolev et un theorem de Fredholme non lineaire pour la transformation conforme de la courbure scalaire*, J. Funct. Anal., 32 (1979), 148–174.

[4] A. Bahri and J. M. Coron; *The Scalar curvature problem on the standard three dimensional sphere*, J. Funct. Anal., 95 (1991), 106–172.

[5] W. Beckner; *Sharp Sobolev inequalities on the sphere and the Moser-Trudinger inequality*, Ann. of Math., 138 (1993), 213–242.

[6] T. Branson; *Differential operators canonically associated to a conformal structure*, Math. Scand. 57 (1985), 293–345.

[7] T. Branson; *Sharp Inequality, the Functional determinant and the Complementary series*, Trans. Amer. Math. Soc., 347 (1995), 3671–3742.

[8] T. Branson, S.-Y. A. Chang and P. Yang; *Estimates and extremals for the zeta-functional determinants on four-manifolds*, Comm. Math. Phys., 149 (1992), no. 2, 241–262.

[9] T. Branson and B. Ørsted; *Explicit functional determinants in four dimensions*, Proc. Amer. Math. Soc., 113 (1991), 669–682.

[10] S. Brendle; *Global existence and convergence for a higher order flow in conformal geometry*, preprint, 2002.

[11] L. A. Caffarelli; *A localization property of viscosity solutions to the Monge-Ampére equation and their strict convexity*, Ann. of Math. (2) 131 (1990), no. 1, 129–134.

[12] L. Caffarelli, L. Nirenberg and J. Spruck; *The Dirichlet problem for nonlinear second order elliptic equations, III: Functions of the eigenvalues of the Hessian*, Acta Math., 155 (1985), no. 3–4, 261–301.

[13] L. Caffarelli and Y. Yang; *Vortex condensation in the Chern-Simons Higgs model: an existence theorem*, Comm. Math. Phys., 168 (1995), 321–336

[14] L. Carleson and S.-Y. A. Chang; *On the existence of an extremal function for an inequality of J. Moser*, Bull. Sci. Math., 110 (1986), 113–127.

[15] K.-C. Chang and J. Q. Liu; *On Nirenberg's problem*, Inter. J. of Math., 4 (1993), 35–58.

[16] S.-Y. A. Chang, M. Gursky and P. Yang; *Prescribing scalar curvature on S^2 and S^3*, Calculus of Variation, 1 (1993), 205–229.

[17] S.-Y. A. Chang, M. Gursky and P. Yang; *Regularity of a fourth order PDE with critical exponent*, Amer. J. of Math., 121 (1999) 215–257.

[18] S.-Y. A. Chang, M. Gursky and P. Yang; *An equation of Monge-Ampere type in conformal geometry, and four-manifolds of positive Ricci curvature*, Ann. of Math. 155 (2002), no. 3, 711–789.

[19] S.-Y. A. Chang, M. Gursky and P. Yang; *An a prior estimate for a fully nonlinear equation on Four-manifolds*, J. D'Analyse Math.; Thomas Wolff memorial issue, 87 (2002), to appear .

[20] S.-Y. A. Chang, M. Gursky and P. Yang; *Entire solutions of a fully nonlinear equation*, preprint 2001.

[21] S.-Y. A. Chang, M. Gursky and P. Yang; *A conformally invariant sphere theorem in four dimensions*, preprint 2002.

[22] S.-Y. A. Chang and J. Qing; *The Zeta functional determinants on manifolds with boundary I—the formula*, J. Funct. Anal., 147 (1997), no.2, 327–362.

[23] S.-Y. A. Chang and J. Qing; *The Zeta functional determinants on manifolds with boundary II—Extremum metrics and compactness of isospectral set*, J. Funct. Anal., 147 (1997), no.2, 363–399.

[24] S.-Y. A. Chang, J. Qing and P. Yang; *On the Chern-Gauss-Bonnet integral for conformal metrics on R^4*, Duke Math J, 103, No. 3 (2000), 523–544.

[25] S.-Y. A. Chang, J. Qing and P. Yang; *Compactification for a class of conformally flat 4-manifold*, Inventiones Mathemticae, 142 (2000), 65–93.

[26] S.-Y. A. Chang and P. Yang; *Prescribing Gaussian curvature on S^2*, Acta Math., 159 (1987), 215–259.

[27] S.-Y. A. Chang and P. Yang; *Conformal deformation of metrics on S^2*, J. Diff. Geom., 27 (1988), no.2, 259–296.

[28] S.-Y. A. Chang and P. Yang; *Extremal metrics of zeta functional determinants on 4-Manifolds*, Ann. of Math., 142 (1995), 171–212.

[29] S.-Y. A. Chang and P. Yang; *On uniqueness of solution of an n-th order differential equation in conformal geometry*, Math. Res Letter, 4 (1997), 91–102.

[30] S.-Y. A. Chang and P. Yang; *The inequality of Moser and Trudinger and applications to conformal geometry*, to appear in Comm. Pure Appl. Math., special issue in memory of Jürgen Moser.

[31] C.-C. Chen and C.-S. Lin; *Topological degree for a mean field equation on Riemann surfaces*, preprint 2001.

[32] C.-C. Chen and C.-S. Lin; *Prescribing Scalar curvature on S^n, Part I, Apriori estimates*, J. Diff. Geom., 57 (2002), 67–171.
[33] S. Cohn-Vossen; *Kürzest Wege und Totalkrümmung auf Flächen*, Compositio Math. 2 (1935), 69–133.
[34] Z. Djadli, E. Hebey and M. Ledoux; *Paneitz-type operators and applications*, Duke Math. J. 104 (2000), no. 1, 129–169
[35] Z. Djadli, A. Malchiodi and M.O. Ahmedou; *Prescribing a fourth order conformal invariant on the standard sphere, Part II: blowup analysis and applications*, preprint 2001.
[36] W.-Y. Ding, J. Jost, J. Li and G. Wang; *Multiplicity results for the two-vertex Chern-Simons Higgs model on the two-sphere*, Comm. Math. Helv., 74 (1999), no. 1, 118–142.
[37] M. Eastwood and M. Singer; *A conformally invariant Maxwell gauge*, Phys. Lett. 107A (1985), 73–83.
[38] C. Evans; *Classical solutions of fully non-linear, convex, second order elliptic equations*, Comm. Pure Appl. Math., XXV (1982), 333–363.
[39] H. Fang; Thesis, Princeton Univ., 2001.
[40] L. Fontana; *Sharp borderline Sobolev inequalities on compact Riemannian manifolds*, Comm. Math. Helv., 68 (1993), 415–454.
[41] C. Fefferman and C. R. Graham; *Conformal invariants*, In: *Élie Cartan et les Mathématiques d'aujourd'hui. Asterisque* (1985), 95–116.
[42] C. Fefferman and C. R. Graham; *Q-curvature and Poincare metrics*, Math. Res. Letters, 9 (2002), no. 2 and 3, 139–152.
[43] M. Flücher; *Extremal functions for the Trudinger-Moser inequality in 2 dimensions*, Comm. Math. Helv., 67 (1992), 471–497.
[44] L. Garding; *An inequality for hyperbolic polynomials*, J. Math. Mech., 8 (1959), 957–965.
[45] C. R. Graham, R. Jenne, L. Mason, and G. Sparling; *Conformally invariant powers of the Laplacian, I: existence*, J. London Math. Soc., 46 (1992), no. 2, 557–565.
[46] C.R. Graham and M. Zworski; *Scattering matrix in conformal geometry*, preprint, 2001.
[47] B. Guan and J. Spruck; *Boundary value problems on S^n for surfaces of constant Gauss curvature*, Ann. of Math. 138 (1993), 601–624.
[48] P. Guan, J. Viaclvosky and G. Wang; *Some properties of the Schouten tensor and applications to conformal geometry*, preprint, 2002.
[49] P. Guan and G. Wang; *A Fully nonlinear conformal flow on locally conformally flat manifolds*, preprint 2002.
[50] M. Gursky; *The Weyl functional, deRham cohomology and Kahler-Einstein metrics*, Ann. of Math., 148 (1998), 315–337.
[51] M. Gursky; *The principal eigenvalue of a conformally invariant differential operator, with an application to semilinear elliptic PDE*, Comm. Math. Phys., 207 (1999), 131–143.
[52] M. Gursky and J. Viaclovsky; *A new variational characterization of three-dimensional space forms*, Invent. Math., 145 (2001), 251–278.

[53] R. Hamilton; *Four manifolds with positive curvature operator*, J. Diff. Geom., 24 (1986), 153–179.
[54] Z.-C. Han; *Prescribing Gaussian curvature on S^2*, Duke Math. J., 61 (1990), 679–703.
[55] C. W. Hong; *A best constant and the Gaussian curvature*, Proc. Amer. Math. Soc., 97 (1986), 737–747.
[56] A. Huber; *On subharmonic functions and differential geometry in the large*, Comm. Math. Helv., 32 (1957), 13–72.
[57] J. Kazdan and F. Warner; *Existence and conformal deformation of metrics with prescribed Gaussian and scalar curvature*, Ann. of Math., 101 (1975), 317–331.
[58] N.V. Krylov; *Boundedly nonhomogeneous elliptic and parabolic equations*, Izv. Akad. Nak. SSSR Ser. Mat., 46 (1982), 487–523; English transl. in Math. USSR Izv., 20 (1983), 459–492.
[59] Y. Li; *Prescribing Scalar curvature on S^n and related problems, Part II: existence and compactness*, Comm. Pure and Appl. Math. XLIX (1996), 541–587.
[60] A. Li and Yanyan Li; *On some conformally invariant fully nonlinear equations*, preprint, 2002.
[61] C. Margerin; *A sharp characterization of the smooth 4-sphere in curvature forms*; CAG, 6 (1998), no. 1, 21–65.
[62] J. Moser; *A Sharp form of an inequality by N. Trudinger*, Indiana Math. J., 20 (1971), 1077–1091.
[63] J. Moser; *On a non-linear problem in differential geometry*, Dynamical Systems, (Proc. Sympos. Univ. Bahia, Salvador, 1971) 273–280. Academic Press, New York 1973.
[64] E. Onofri; *On the positivity of the effective action in a theory of random surfaces*, Comm. Math. Phys., 86 (1982), 321–326.
[65] B. Osgood, R. Phillips, and P. Sarnak; *Extremals of determinants of Laplacians*, J. Funct. Anal., 80 (1988), 148–211.
[66] B. Osgood, R. Phillips, and P. Sarnak; *Compact isospectral sets of surfaces*, J. Funct. Anal., 80 (1988), 212–234.
[67] K. Okikiolu: *The Campbell-Hausdorff theorem for elliptic operators and a related trace formula*, Duke Math. J., 79 (1995) 687–722.
[68] K. Okikiolu; *Critical metrics for the determinant of the Laplacian in odd dimensions*, Ann. of Math., 153 (2001), no. 2, 471–531.
[69] A.V. Pogorelev; *The Dirichlet problem for the multidimensional analogue of the Monge-Ampere equation*, Dokl. Acad. Nank. SSSR, 201(1971), 790–793. In English translation, Soviet Math. Dokl. 12 (1971), 1227–1231.
[70] S. Paneitz; *A quartic conformally covariant differential operator for arbitrary pseudo-Riemannian manifolds*, Preprint, 1983.
[71] A. Polyakov; *Quantum geometry of Bosonic strings*, Phys. Lett. B, 103 (1981), 207–210.
[72] K. Richardson; *Critical points of the determinant of the Laplace operator*, J. Funct. Anal., 122 (1994), 52–83.
[73] D. B. Ray and I. M. Singer; *R-torsion and the Laplacian on Riemannian manifolds*, Advances in Math., 7 (1971), 145–210.

[74] R. Schoen, *Conformal deformation of a Riemannian metric to constant scalar curvature*, J. Diff. Geom., 20 (1984), 479–495.

[75] R. Schoen and D. Zhang; *Prescribing scalar curvature on S^n*, Calculus of Variations, 4 (1996), 1–25.

[76] R. Schoen and S.T. Yau; *Conformally flat manifolds, Kleinian groups and scalar curvature*, Invent. Math. 92 (1988), 47–71.

[77] T.-L. Soong; *Extremal functions for the Moser inequality on S^2 and S^4*, Ph.D Thesis, UCLA 1991.

[78] M. Struwe and G. Tarantello; *'On multi-vortex solutions in Chern-Simon Gauge theory*, Boll. Unione Math. Ital. Sez. B Artic. Mat., (8) 1 (1998), 109–121.

[79] G. Tarantello; *Multiple condensate solutions for the Chern-Simons-Higgs theory*, J. Math. Phys., 37 (1996), 3769–3796.

[80] N. Trudinger; *On embedding into Orlicz spaces and some applications*, J. Math. Mech., 17 (1967), 473–483.

[81] N. Trudinger; *Remarks concerning the conformal deformation of Riemannian structure on compact manifolds*, Ann. Scuolo Norm. Sip. Pisa, 22 (1968), 265–274.

[82] K. Uhlenbeck and J. Viaclovsky; *Regularity of weak solutions to critical exponent variational equations*, Math. Res. Lett., 7 (2000), 651–656.

[83] J. Viaclovsky; *Conformal geometry, Contact geometry and the Calculus of Variations*, Duke Math. J., 101 (2000), no.2, 283–316.

[84] X.-W. Xu and P. Yang; *On a fourth order equation in 3-D*, to appear in ESAIM Control Optim. Calc. Var.

[85] H. Yamabe; *On a deformation of Riemannian structures on compact manifolds*, Osaka Math. J., 12 (1960), 21–37.

ICM 2002 · Vol. I · 209–233

Emerging Applications of Geometric Multiscale Analysis

David L. Donoho*

Abstract

Classical multiscale analysis based on wavelets has a number of successful applications, e.g. in data compression, fast algorithms, and noise removal. Wavelets, however, are adapted to point singularities, and many phenomena in several variables exhibit intermediate-dimensional singularities, such as edges, filaments, and sheets. This suggests that in higher dimensions, wavelets ought to be replaced in certain applications by multiscale analysis adapted to intermediate-dimensional singularities,

My lecture described various initial attempts in this direction. In particular, I discussed two approaches to geometric multiscale analysis originally arising in the work of Harmonic Analysts Hart Smith and Peter Jones (and others): (a) a directional wavelet transform based on parabolic dilations; and (b) analysis via anistropic strips. Perhaps surprisingly, these tools have potential applications in data compression, inverse problems, noise removal, and signal detection; applied mathematicians, statisticians, and engineers are eagerly pursuing these leads.

Note: Owing to space constraints, the article is a severely compressed version of the talk. An extended version of this article, with figures used in the presentation, is available online at:

http://www-stat.stanford.edu/~donoho/Lectures/ICM2002

2000 Mathematics Subject Classification: 41A30, 41A58, 41A63, 62G07, 62G08, 94A08, 94A11, 94A12, 94A29.
Keywords and Phrases: Harmonic analysis, Multiscale analysis, Wavelets, Ridgelets, Curvelets, Directional wavelets.

1. Prologue

Since the last ICM, we have lost three great mathematical scientists of the twentieth century: Alberto Pedro Calderón (1922-1999), John Wilder Tukey (1915-2000) and Claude Elwood Shannon (1916-2001). Although these three are not

*Department of Statistics, Stanford University, Stanford, CA 94305, USA. E-mail: donoho@stanford.edu

typically spoken of as a group, I find it fitting to mention these three together because each of these figures symbolizes for me one aspect of the *unreasonable effectiveness of harmonic analysis.*

Indeed we are all aware of the birth of harmonic analysis in the nineteenth century as a tool for understanding of the equations of mathematical physics, but it is striking how the original tools of harmonic analysis have frequently (a) changed, and (b) been applied in ways the inventors could not have anticipated. Thus, (a) harmonic analysis no longer means 'Fourier Analysis' exclusively, because wavelet and other forms of decompositions have been invented by modern harmonic analysts (such as Calderón); and (b) harmonic analysis finds extensive application outside of mathematical physics, as a central infrastructural element of the modern information society, because of the ubiquitous applications of the fast Fourier transform (after Tukey) and Fourier transform coding (after Shannon).

There is a paradox here, because harmonic analysts are for the most part not seeking applications, or at any rate, what they regard as possible applications seem not to be the large-scale applications that actually result. Hence the impact achieved by harmonic analysis has often not been the intended one After meditating for a while on what seems to be the 'unreasonable' effectiveness of harmonic analysis, I have identified what seems to me a chain of argumentation that renders the 'unreasonable' at least 'plausible'. The chain has two propositions:

- *Information has its own architecture.* Each data source, whether imagery, sound, text, has an inner architecture which we should attempt to discover and exploit for applications such as noise removal, signal recovery, data compression, and fast computation.
- *Harmonic Analysis is about inventing and exploring architectures for information.* Harmonic analysts have always created new architectures for decomposition, rearrangement and reconstruction of operators and functions.

In short, *the inventory of architectures created by harmonic analysis amounts to an* **intellectual patrimony** *which modern scientists and engineers can fruitfully draw upon for inspiration as they pursue applications.* Although there is no necessary connection between the architectures that harmonic analysts are studying and the architectures that information requires, it is important that we have many examples of useful architectures available, and harmonic analysis provides many of these. Occasionally, the architectures already inventoried by harmonic analysts will be exactly the right ones needed for specific applications.

I stress that the 'externally professed goals' of harmonic analysis in recent decades have always been theorems, e.g. about the almost everywhere convergence of Fourier Series, the boundedness of Bochner-Riesz summation operators, or the boundedness of the Cauchy integral on chord-arc curves. These externally professed goals have, as far as I know, very little to do with applications where harmonic analysis has had wide scale impact. Nevertheless, some harmonic analysts are aware of the architectural element in what they do, and value it highly. As R.R. Coifman has pointed out to me in private communication:

> "The objective of Zygmund, Calderón and their school was not the establishment of new theorems by any means possible. It was often to

> take known results that seemed like magic — e.g. because of the way they used complex variables methods — and tear them apart, finding the underlying structures and their inner interactions that made it absolutely clear what was going on. The test of understanding was measured by the ability to prove an estimate."

In short, the goal was to find the right architecture, not merely to find the right estimate.

2. Overview

In my lecture, I was able to discuss the possibility that a coherent subject of *Geometric* Multiscale Analysis (GMA) can be developed – a subject spanning both mathematics and a wide range of applications. It is at this point unclear what the boundaries of the subject will be, but perhaps the speculative nature of what I had to say will excite the interest of some readers. I found it useful to organize the presentation around the Calderón reproducing formula, which gave us the continuous wavelet transform, but also can be adapted to give us other multiscale transforms with interesting geometric aspects. The several different information architectures I described give an intuitive understanding of what GMA might consist of. In the article below, I will review some of the achievements of classical 1-dimensional multiscale analysis (wavelet analysis) starting in the 1980's, both the mathematical achievements and the extensive applications; then I will as a warm-up discuss reasons that we need alternatives to 1-dimensional multiscale analysis and its straightforward d-dimensional extensions, and some ideas such as ridgelets, that point in the expected directions. In my lecture, I was able to discuss two harmonic analysis results of the 1990's – Hart Smith's "Hardy space for FIO's" and Peter Jones' "Travelling Salesman" theorem. Both results concern the higher-dimensional setting, where it becomes possible to bring in geometric ideas. I suggested that, in higher dimensions, there are interesting, nontrivial, nonclassical, geometric multiscale architectures, with applications paralleling the one-dimensional case. I was able to sketch some developing applications of these post-classical architectures. If these applications can be developed as extensively as has been done for classical multiscale analysis, the impacts may be large indeed. In this article, I really have space only to mention topics growing out of my discussion of Hart Smith's paper. For an extended version of the article, covering the talk more fully, see [26].

Note: Below we make a distinction between stylized applications (idealized applications in mathematical models) and actual applications (specific contributions to scientific discourse and technological progress); we always describe the two in separate subsections.

3. Classical multiscale analysis

An efficient way to introduce classical multiscale analysis is to start from Calderón's reproducing formula, or as commonly called today, the *Continuous*

Wavelet Transform. We suppose we have a real-valued function $f : \mathbf{R} \mapsto \mathbf{R}$ which we want to decompose into contributions from various scales and locations. We take with a **wavelet**, an oscillatory real-valued function $\psi(t)$ satisfying the Calderón admissibility condition imposed on the Fourier transform $\hat{\psi}$ as $\int_0^\infty |\hat{\psi}(\xi t)|^2 \frac{dt}{t} = 2\pi$, $\forall \xi \neq 0$. We translate and dilate according to $(\psi_{a,b})(t) = \psi((t-b)/a)/\sqrt{a}$. We perform **Wavelet Analysis** by 'hitting' the function against all the different wavelets, obtaining $W_f(a,b) = \langle \psi_{a,b}, f \rangle$; W_f is called the Continuous Wavelet Transform (CWT). The CWT contains all the information necessary to reconstruct f, so we can perform **Wavelet Synthesis** by integrating overall all scales and locations, summing up wavelets with appropriate coefficients.

$$f(t) = \int W_f(a,b)\psi_{a,b}(t)\mu(dadb).$$

Here $\mu(dadb)$ is the appropriate reference measure, in this case $\frac{db}{a}\frac{da}{a}$. The 'tightness' of the wavelet transform as a characterisation of the properties of f is expressed by the Parseval-type relation $\int W_f(a,b)^2 \mu(da\, db) = \int f(t)^2 dt$. See also [16, 36, 42].

3.1. Mathematical results

The CWT maps f into a time-scale plane; by measuring properties of this time-scale portrait we can obtain norms on functions which lead to interesting theories of functional spaces and their properties; there are two broad scales of such spaces we can describe. To define the **Besov** $B^\sigma_{p,q}$ spaces we integrate over locations first, and then over scales

$$\left(\int \left(\int (|W(a,b)|a^{-s})^p \frac{db}{a} \right)^{q/p} \frac{da}{a} \right)^{1/p}.$$

To define the **Triebel-Lizorkin** $F^\sigma_{p,q}$ spaces we integrate over scales first and then over locations

$$\left(\int \left(\int (|W(a,b)|a^{-s})^q \frac{da}{a^{1+q/p}} \right)^{p/q} db \right)^{1/p}.$$

Here $s = \sigma - (1/p - 1/2)$, and we adopt a convention here and below of *ignoring the low frequencies* so that actually these formulas are only correct for functions which are built from frequencies $|\xi| > \lambda_0$; the correct general formulas would require an extra term for the low frequencies which will confuse the novice and be tedious for experts. Also for certain combinations of parameters $p, q = 1, \infty$ for example, changes ought to be made, based on maximal functions, BMO norms, etc., but in this expository work we gloss over such issues.

Each of these norms asks that the wavelet transform decay as we go to finer scales, and so controls the oscillations of the functions. Intuition about some of these spaces comes by thinking of a wavelet coefficient as something akin to a difference operator such as $f(b+a) - 2f(b) + f(b-a)$; the various norms on the continuous wavelet coefficients measure explicitly the finite differences and implicitly

the derivatives of the analyzed functions. The distinctions between spaces come in the subtle aspects of choice of order of integrating in scale and in location and in choice of p and q. We get the following sequence of relations between the spaces defined by the F and B scales and classical spaces:

- L^p : $L^p \sim F^0_{p,2}$, $1 < p < \infty$.
- H^p : $H^p \sim F^0_{p,2}$, $0 < p \leq 1$.
- **Sobolev:** $W^m_p \sim F^m_{p,2}$, $1 < p < \infty$.
- **Hölder:** $C^\alpha \sim B^\alpha_{\infty,\infty}$

There are also equivalences with non-classical, but very interesting, spaces, such as the Bump Algebra $B^1_{1,1}$, and almost-equivalences to some other fundamental spaces, such as $BV(\mathbf{R})$. The full story about such equivalence is told very well in [42, 36].

An important structural fact about these spaces is that they admit molecular decompositions; we can define molecules as functions obeying certain size, smoothness and vanishing moment conditions, which are localized near an interval of some scale and location, and then show that, although elements of these spaces are defined by norms on the continuum domain, functions belong to these spaces if and only if they can be written as superpositions $f(x) = \sum_Q A_Q m_Q(x)$ where m_Q are molecules and the A_Q are scalar coefficients, and where the coefficient sequence $(A_Q)_Q$ obeys certain norm constraints. Results of this kind first emerged in the 1970's; a canonical way to get such results uses the CWT [36]. Consider the dyadic cells

$$Q = \{(a,b) : 2^{-j} > a \geq 2^{-(j+1)}, k/2^j \leq b < (k+1)/2^j\},$$

note that they obey $\mu(Q) \approx 1$; they are "unit cells' for the reference measure. It turns out that the behavior of $W(a,b)$ at various points within such a cell Q stays roughly comparable [16, 36], and that the $\psi_{a,b}$ all behave similarly as well. As a result, the integral decomposition offered by the Calderón reproducing formula can sensibly be discretized into terms arising from different cells.

$$\begin{aligned} f(x) &= \int W(a,b)\psi_{a,b}(x)\mu(da\,db) \\ &= \sum_Q \int_Q W(a,b)\psi_{a,b}(x)\mu(da\,db) \\ &= \sum_Q M_Q(x), \qquad M_Q = \int_Q W(a,b)\psi_{a,b}\mu(da\,db) \\ &= \sum_Q A_Q m_Q(x), \qquad A_Q = \|W(\cdot,\cdot)\|_{L^2(Q)} \end{aligned}$$

Now roughly speaking, each m_Q is a mixture of wavelets at about the same location and scale, and so is something like a wavelet, a coherent oscillatory waveform of a certain location and scale. This type of discretization of the Calderón reproducing formula has been practiced since the 1970's, for example by Calderón, and by Coifman and Weiss [13, 14], who introduced the terms *molecular decomposition* (and atomic decomposition) for discrete series of terms localized near a certain scale and

location. Hence, the m_Q may be called molecules and the A_Q represent the contributions of various molecules. The spaces $F^\sigma_{p,q}$ and $B^\sigma_{p,q}$ can then be characterized by the decomposition $f(x) = \sum_Q A_Q m_Q(x)$: we can define sequence-space norms $f^\sigma_{p,q}$ as in (3.2) below for which

$$\|f\|_{F^\sigma_{p,q}} \sim \|(A_Q)_Q\|_{f^\sigma_{p,q}},$$

and similarly for Besov sequence norms $b^\sigma_{p,q}$. This gives a clear understanding of the structure of f in terms of the distribution of the number and size of oscillations across scales.

While the molecular decomposition is very insightful and useful for proving structure theorems about functional spaces, it has two drawbacks which severely restrict practical applications. First, the A_Q are nonlinear functionals of the underlying object f; secondly, the m_Q are variable objects which depend on f. As a result, practical applications of the sum $\sum_Q A_Q m_Q$ are not as straightforward as one might like. Starting in the early 1980's, it was found that a much simpler and more practical decomposition was possible; in fact with appropriate choice of generating wavelet – different than usually made in the CWT – one could have an *orthonormal wavelet basis* [42, 16].

$$f = \sum_{j,k} W(2^{-j}, k/2^j)\psi_{2^{-j},k/2^j} = \sum_{j,k} \alpha_{j,k}\psi_{j,k}. \tag{3.1}$$

Essentially, instead of integrating over dyadic cells Q, it is necessary only to sample once per cell! Several crucial advantages in applications flow from the fact that the coefficients $\alpha_{j,k}$ are linear in f and the $\psi_{j,k}$ are fixed and known.

A theoretical advantage flows from the fact that the same norm equivalence that was available for the amplitudes (A_Q) in the molecular decomposition also applies for the wavelet coefficients:

$$\|f\|_{B^\sigma_{p,q}} \sim \|\alpha\|_{b^\sigma_{p,q}} \equiv \left(\sum_j (\sum_k |\alpha_{j,k}|^p)^{q/p} 2^{jsq} \right)^{1/q}, \tag{3.2}$$

$$\|f\|_{F^\sigma_{p,q}} \sim \|\alpha\|_{f^\sigma_{p,q}} \equiv \left(\int (\sum_j |\alpha_{j,k}|^q 2^{jsq} \chi_{j,k}(t))^{p/q} \right)^{1/q}. \tag{3.3}$$

This implies that the wavelets $\psi_{j,k}$ make an unconditional basis for appropriate spaces in the Besov and Triebel scales. This can be seen from the fact that the norm involves only $|\alpha_{j,k}|$; a fact which is quite different from the case with Fourier analysis. Unconditionality implies that the balls $\{f : \|f\|_{B^\sigma_{p,q}} \le A\}$ are closely inscribed by and circumscribed by balls $\{f : \|\alpha\|_{b^\sigma_{p,q}} \le A'\}$ which are quite simple geometric objects, solid and orthosymmetric with respect to the wavelets as 'principal axes'. This solid orthosymmetry is of central significance for the optimality of wavelets for many of the stylized applications mentioned below; compare [20, 22, 32].

Our last chapter in the mathematical development of classical multiscale methods concerns the connection between Besov spaces and approximation spaces. In the

late 1960's, Jaak Peetre observed that the space $B^{\sigma}_{1/\sigma,1/\sigma}$, $\sigma > 1$ was very special. It served as the *Approximation Space* for approximation in L^{∞} norm by free knot splines, i.e. as the set of functions approximable at rate $n^{-\sigma}$ by splines with n free knots. In the 1980's a more general picture emerged, through work of e.g. Brudnyi, DeVore, Popov and Peller [18]: that the space $B^{\sigma}_{\tau,\tau}$ served as the approximation space of many nonlinear approximation schemes (e.g. rational functions), under L^p approximation error, where $1/\tau = \sigma + 1/p$. This says that although $\tau < 1$ at first seems unnatural (because graduate mathematical training emphasizes convex spaces) these nonconvex spaces are fundamental. The key structural fact is that those spaces are equivalent, up to renorming, to the set of functions whose wavelet coefficients belong to an ℓ^{τ} ball, $\tau < 1$. Hence, membership of wavelet coefficients in an ℓ^{τ} ball for small τ becomes of substantial interest. The intuitive appeal for considering ℓ^{τ} balls is clear by considering the closely-related weak-ℓ^{τ} balls; they can be defined as the constants C in relations of the form

$$\mu\{(a,b) : |W(a,b)| > \epsilon\} \leq C\epsilon^{-1/\tau}, \quad \epsilon > 0,$$

or

$$\#\{(j,k) : |\alpha_{j,k}| > \epsilon\} \leq C\epsilon^{-1/\tau}, \quad \epsilon > 0.$$

They are visibly measures of the sparsity in the time-scale plane, and hence sparsity of that plane controls the asymptotic behavior of numerous nonlinear approximation schemes.

3.2. Stylized applications

We now mention some stylized applications of classical multiscale thinking, i.e. applications in a model world where we can prove theorems in the model setting.

3.2.1. Nonlinear approximation

Since the work of D.J. Newman in the 1960's it was understood that approximation by rational functions could be dramatically better than approximation by polynomials; for example the absolute value function $|t|$ on the interval $[-1,1]$ can be approximated at an exponential rate in n by rational functions with numerator and denominator of degree n, while it can be approximated only at an algebraic rate n^{-1} by polynomials of degree n. While this suggests the power of rational approximation, it must also be noted that rational approximation is a highly nonlinear and computationally complex process.

On the other hand, from the facts (a) that wavelets provide an unconditional basis for Besov spaces, and (b) that certain Besov spaces are approximation spaces for rational approximation, we see that wavelets give an effective algorithm for the same problems where rational functions would be useful. Indeed, based on work by DeVore, Popov, Jawerth, Lucier we know that if we consider the class of functions approximable at rate $\approx n^{-\tau}$ by rational approximation, these same functions can be approximated at the same rate simply by taking a partial reconstruction based on the n "biggest" wavelet coefficients.

In short, from the viewpoint of asymptotic rates of convergence, thresholding of wavelet coefficients - a very weakly nonlinear approximation scheme - is fully as effective as best rational approximation. The same assertion can be made comparing nonlinear approximation by wavelets and by free knot splines.

3.2.2. Data compression

Consider the following mathematical idealization of data compression. We have a function which is an unspecified element of a Besov Ball $\mathcal{F} = \{f : \|f\|_{B^\sigma_{p,q}} \leq A\}$ and we wish to have a coder/decoder pair which can approximate any such function to within an ϵ-distance in L^2 norm by encoding into, and decoding from, a finite bitstring.

In mathematical terms, we are studying the Kolmogorov ϵ-entropy: we wish to achieve $N(\epsilon, \mathcal{F})$, the minimal number of bits required to represent every f in $\mathcal{F}$ to within an L^2 error ϵ. This is known, since Kolmogorov and Tikhomirov, to behave as

$$N(\epsilon, \mathcal{F}) \asymp \epsilon^{-1/\sigma}, \epsilon \to 0. \tag{3.4}$$

Now, up to renorming, the ball $\mathcal{F}$ is isometric to a ball in sequence space $\Theta = \{\alpha : \|\alpha\|_{b^\sigma_{p,q}} \leq A\}$. Such a ball is a subset of $w\ell^\tau$ for $1/\tau = \sigma + 1/2$ and each element in it can be approximated in ℓ^2 error at a rate $M^{-1/\tau+1/2}$ by sparse vectors containing only M nonzero coefficients. Here is a simple coder inspired by this fact. Pick $M(\epsilon)$ coefficients such that the ℓ^2-error of such an approximation is at most (say) $\epsilon/2$. The $M(\epsilon)$ coefficients achieving this can be quantized into integer multiples of a base quantum q, according to $a_{j,k} = \lfloor \alpha_{j,k}/q \rfloor$, with the quantum chosen so that the quantized vector $\alpha^{(q)}$ defined by $\alpha^{(q)}_{j,k} = q \cdot a_{j,k}$, approximates the original coefficients to within ℓ^2 error $\epsilon/2$. The resulting integers $a_{j,k}$ represent the function f to within L^2 error ϵ and their indices can be coded into bit strings, for a total encoding length of not worse that $O(\log(\epsilon^{-1})M(\epsilon)) = O(\log(\epsilon^{-1})\epsilon^{-1/\sigma})$. Hence a very simple algorithm on the wavelet coefficients gets close to the optimal asymptotics (3.4)! Underlying this fact is the geometry of the body Θ; because of its solid orthosymmetry, it contains many high-dimensional hypercubes of ample radius. Such hypercubes are essentially incompressible.

In fact the $\log(\epsilon^{-1})$ factor is removable in a wide range of σ, p, q. In many cases, the ϵ-entropy can be attained, within a constant factor, by appropriate level-dependent scalar quantization of the wavelet coefficients followed by run-length encoding. In other work, Cohen et al. have shown that by using the tree-organization of wavelet coefficients one can develop algorithms which give the right order of asymptotic behavior for the across many smoothness classes; e.g. [12].

In fact more is true. Suppose we use for Besov ball simply the ball $\{f : \|\alpha(f)\|_{b^\sigma_{p,q}} \leq A\}$ based on wavelet coefficients; then by transform coding as in [25] we can get efficient codes with codelength precisely asymptotic equivalence to the Kolmogorov ϵ-entropy by levelwise ℓ^p-sphere vector quantization of wavelet coefficients. Underlying this fact, the representation of the underlying functional class as an orthosymmetric body in infinite-dimensional space is very important.

3.2.3. Statistical estimation

Consider the following mathematical idealization of nonparametric curve estimation. We have an unknown function $f(t)$ on $[0,1]$ which is an element of a Besov Ball $\mathcal{F} = \{f : \|f\|_{B^\sigma_{p,q}} \leq A\}$ We observe data Y from the white noise model

$$Y(dt) = f(t)dt + \epsilon W(dt),$$

where the $W(t)$ is a Wiener process and ϵ the noise level, and we wish to reconstruct f accurately. We measure risk using the mean squared error

$$R_\epsilon(f, \hat{f}) = E\|f - \hat{f}\|_2^2.$$

and evaluate quality by the minimax risk

$$\min_{\hat{f}} \max_{f \in \mathcal{F}} R_\epsilon(f, \hat{f}).$$

Over a wide range of σ, p, q, this minimax risk tends to zero at the rate $(\epsilon^2)^{2\sigma/(2\sigma+1)}$.

In this setting, some simple algorithms based on noisy wavelet coefficients $y_{j,k} = \int \psi_{j,k}(d)Y(dt)$ can be quite effective. In effect, $y_{j,k} = \alpha_{j,k} + \epsilon z_{j,k}$, where $z_{j,k}$ is a white Gaussian noise. By simply applying thresholding to the noisy wavelet coefficients of Y,

$$\hat{\alpha}_{j,k} = y_{j,k} 1_{\{|y_{j,k}| > \lambda\epsilon\}}$$

at scales $0 \leq j \leq \log_2(\epsilon^{-2})$ with threshold $\sqrt{2\log(\epsilon^{-1})}$, we obtain a new set of coefficients; using these we obtained a nonlinear approximation $\hat{f} = \sum_{j,k} \hat{\alpha}_{j,k}\psi_{j,k}$. The quantitative properties are surprisingly good; indeed, using again the $w\ell^\tau$ embedding of the Besov body $b^\sigma_{p,q}$, we have that the ℓ^2-error of nonlinear approximation to α using M terms converges at rate $M^{-1/\tau+1/2}$. Heuristically, the coefficients surviving thresholding have errors of size $\approx \epsilon$, and the object can be approximated by at most M of these with ℓ^2 error $\approx M^{-1/\tau+1/2}$; simple calculations suggest that the risk of the estimator is then roughly $\epsilon^2 \cdot M + M^{-2/\tau+1}$ where M is the number of coefficients larger than ϵ in amplitude; this is the same order as the minimax risk $\epsilon^{2\sigma/(2\sigma+1)}$! (Rigorous analysis shows that for this simple algorithm, log terms intervene [31].) If we are willing to refine the thresholding in a level-dependent way, we can obtain a risk which converges to zero at the same rate as the minimax risk as $\epsilon \to 0$, e.g. [32]. Moreover, if we are willing to adopt as our Besov norm the sequence space $b^\sigma_{p,q}$ norm based on wavelet coefficients, then by applying a sequence of particular scalar nonlinearities to the noisy wavelet coefficients (which behave qualitatively like thresholds) we can get precise asymptotic equivalence to the minimax risk, i.e. precise asymptotic minimaxity [32]. Parallel results can be obtained with wavelet methods in various inverse problems, where f is still the estimand, but we observe noisy data on Kf rather than f, with K a linear operator, such as convolution or Radon transform [21].

3.2.4. Fast computation

An important theme for scientific computation is the sparse representation, not of functions, but of operators. For this purpose a central fact pointed out by Yves Meyer [42] is that wavelets sparsify large classes of operators. Let T be a Calderon-Zygmund operator (CZO); the matrix representation of such operator in the wavelet basis

$$M_{j,k}^{i,l} = \langle \psi_{j,k}, T\psi_{i,l} \rangle,$$

then M is sparse – all its rows and columns have finite ℓ^p norms for each $p > 0$. In short, such an operator involves interactions between very few pairs of terms.

For implications of such sparsity, consider the work of Beylkin, Coifman, and Rokhlin [3]. Suppose T is a CZO, and let $Comp(\epsilon, n)$ denote the number of flops required to compute an ϵ-approximation to $P_n T P_n$, where P_n is an projector onto scales larger than $1/n$. In [3] it was shown that, ignoring set-up costs,

$$Comp(\epsilon, n) = O(\log(1/\epsilon)n)$$

so that such operators could be applied many times with cost essentially linear in problem size, as opposed to the $O(n^2)$ cost nominally demanded by matrix multiplication. The algorithm was roughly this: represent the operator in a wavelet basis, threshold the coefficients, and keep the large coefficients in that representation. A banded matrix results, which can be applied in order $O(n)$ flops. (The story is a bit more subtle, since the algorithm as written would suffer an additional $O(\log(n))$ factor; to remove this, Beylkin, Coifman, and Rokhlin's nonstandard form must be applied.)

3.3. Applications

The possibility of applying wavelets to real problems relies heavily on the breakthrough made by Daubechies [15] (building on work of Mallat [41]) which showed that it was possible to define a wavelet transform on finite digital signals which had orthogonality and could be computed in order n flops. Once this algorithm was available, a whole range of associated fast computations followed. Corresponding to each of the 'stylized applications' just listed, many 'real applications' have been developed over the last decade; the most prominent are perhaps the use of wavelets as part of the JPEG-2000 data compression standard, and in a variety of signal compression and noise-removal problems. For reasons of space, we omit details, referring the reader instead to [33] and to various wavelet-related conferences and books.

4. Need for geometric multiscale analysis

The many successes of classical multiscale analysis do not exhaust the opportunities for successful multiscale analysis. The key point is the slogan we formulated earlier - *Information has its own architecture.* In the Information Era, where new

data sources are proliferating endlessly, each with its own peculiarities and specific phenomena, there is a need for expansions uniquely adapted to each type of data.

In this connection, note that classical wavelet analysis is uniquely adapted to objects which are smooth apart from *point singularities.* If a function is C^∞ except for step discontinuities at a finite set of points, its continuous wavelet transform will be very sparse. In consequence, the decreasing rearrangement of its wavelet coefficients will decay rapidly, and n-term approximations to the object will converge rapidly in L^2 norm. With the right definitions the story in high dimensions is similar: wavelets give a sparse representation of point singularities.

On the other hand, for singularities along lines, planes, curves, or surfaces, the story is quite different. For functions in dimension 2 which are discontinuous along a curve, but otherwise smooth, the 2-dimensional CWT will not be sparse. In fact, the the decreasing rearrangement of its wavelet coefficients will decay like C/N, and N-term approximations to the object will converge no faster than $O(N^{-1})$ in squared L^2 norm. Similar statements can be made for singularities of dimension $0 < k < d$ in dimension d. In short, wavelets are excellent for representing smooth data containing point singularities but not singularities of intermediate dimensions.

There are many examples of data where singularities of intermediate dimensions constitute important features. One example comes from extragalactic astronomy, where gravitational clustering has caused matter to congregate in 'filaments' and 'sheets' in 3-dimensions. Another example comes from image analysis, say of SAR imagery, where stream beds, ridge lines, roads and other curvilinear phenomena punctuate the underlying background texture. Finally, recently-developed tools for $3D$ imaging offer volumetric data of phsyical objects (eg biological organs) where sheetlike structures are important.

We can summarize our vision for the future of multiscale analysis as follows.

If it is possible to sparsely analyze objects which are smooth apart from intermediate-dimensional singularities, this may open **new vistas in mathematical analysis**, *offering (a) new functional Spaces, and (b) new representation of mathematically important operators.*

If, further, it is possible algorithmize such analysis tools, this would open new applications involving (a) data compression; (b) noise removal and recovery from Ill-posed inverse problems; (c) feature extraction and pattern recognition; and (d) fast solution of differential and integral equations.

But can we realistically expect to sparsely analyse such singularities? By considering Calderón-like formulas, we can develop some understanding.

4.1. Ridgelet analysis

We consider first the case of singularities of co-dimension 1. It turns out that the ridgelet transform is adapted to such singularities.

Starting from an admissible wavelet ψ, define the ridgelet $\rho_{a,b,\theta}(x) = \psi_{a,b}(u'_\theta x)$, where u_θ is a unit vector pointing in direction θ and so this is a wavelet in one direction and constant in orthogonal directions [6]. In analogy to the continuous wavelet transform, define the continuous ridgelet transform $R_f(a, b, \theta) = \langle \rho_{a,b,\theta}, f \rangle$. There

is a synthesis formula

$$f(x) = \int R_f(a,b,\theta)\rho_{a,b,\theta}(x)\mu(da\,db\,d\theta)$$

and a Parseval relation

$$\|f\|_2^2 = \int R_f(a,b,\theta)^2\mu(da\,db\,d\theta)$$

both valid for an appropriate reference measure μ. Note the similarity to the Calderón formula.

In effect this is an analysis of f into contributions from 'fat planes'; it has been extensively developed in Emmanuel Candès' Stanford thesis (1998) and later publications. Suppose we use it to analyze a function $f(x) \in L^2(\mathbf{R}^n)$ which is smooth apart from a singularity across a hyperplane. If our function is, say, $f_{u,a}(x) = 1_{\{u'x>a\}}e^{-\|x\|^2}$, Candès [5]. showed that the ridgelet transform of $f_{u,a}$ is sparse. For example, a sampling of the continuous ridgelet transform at dyadic locations and scales and directions gives a set of coefficients such that the rearranged ridgelet coefficients decay rapidly. It even turns out that we can define "orthonormal ridgelets" (which are not true ridge functions) such that the orthonormal ridgelet coefficients are sparse: they belong to every ℓ^p with $p > 0$ [24]. In short, an appropriate multiscale analysis (but not wavelet analysis) successfully compresses singularities of co-dimension one.

4.2. *k*-plane ridgelet transforms

We can develop comparable reproducing formulas of co-dimension k in R^d. If P_k denotes orthoprojector onto a k-plane in R^d, and ψ an admissible wavelet for k-dimensional space, we can define a k-plane Ridgelet: $\rho_{a,b,P_k}(x) = \psi_{a,b}(P_k x)$ and obtain a k-plane ridgelet analysis: $R_f(a,b,P_k) = \langle \rho_{a,b,P_k}, f\rangle$. We also obtain a reproducing formula

$$f(x) = \int R_f(a,b,P_k)\rho_{a,b,P_k}(x)\mu(da\,db\,dP_k)$$

and a Parseval relation $\|f\|_2^2 = \int R_f(a,b,P_k)^2\mu(da\,db\,dP_k)$, with in both cases $\mu()$ the appropriate reference measure. In short we are analyzing the object f into 'Fat Lines', 'Fat k-planes,' $1 \le k \le n-1$. Compare [23]. Unfortunately, all such representations have drawbacks, since to use them one must fix in advance the co-dimension k; moreover, very few singularities are globally flat!

4.3. Wavelet transforms for the full affine group

A more ambitious approach is to consider wavelets indexed by the general affine group $GA(n)$; defining $(\psi_{A,b}g)(x) = \psi(Ax+b)\cdot|A|^{1/2}$. This leads to the wavelet analysis $W_f(A,b) = \langle \psi_{A,b}, f\rangle$. Taking into account the wide range of'anisotropic

dilations and directional preferences poossible within such a scheme, we are analyzing f by waveforms which represent a very wide range of behaviors: 'Fat Points', 'Fat Line Segments', 'Fat Patches', and so on.

This exciting concept unfortunately fails. No matter what wavelet we pick to begin with, $\int W_f(A,b)^2 \mu(dAdb) = +\infty$. (technically speaking, we cannot get a square-integrable representation of the general affine group; the group is too large) [46, 47]. Moreover, synthesis fails: $\int W_f(A,b)\psi_{a,b}(t)\mu(dA\,db)$ is not well-defined. Finally, the transform is not sparse on singularities.

In short, the dream of using Calderón-type formulas to easily get a decomposition of piecewise smooth objects into 'Fat Points', 'Fat Line Segments', 'Fat Surface Patches', and so on fails. Success will require hard work.

4.4. A cultural lesson

The failure of soft analysis is not unexpected, and not catastrophic. As Jerzy Neyman once said: life is complicated, but not uninteresting. As Lennart Carleson said:

> There was a period, in the 1940's and 1950's, when classical analysis was considered dead and the hope for the future of analysis was considered to be in the abstract branches, specializing in generalization. As is now apparent, the death of classical analysis was greatly exaggerated ... the reasons for this ... [include] ... the realization that in many problems complications cannot be avoided, and that intricate combinatorial arguments rather than polished theories are in the center.

Our response to the failure of Calderón's formula for the full $Ax+b$ group was to consider, in the ICM Lecture, two specific strategies for decomposing multidimensional objects. In the coming section, we will consider analysis using a special subset of the $Ax+b$ group, where a Calderón-like formula still applies, and we can construct a fairly complete analog of the wavelet transform – only one which is efficient for singularities of co-dimension 1. In the lecture (but not in this article), we also considered analysis using a fairly full subset of the $Ax+b$ group, but in a simplified way, and extracted the results we need by special strategies (viz. Carleson's "intricate combinatorial arguments") rather than smooth general machinery. The results delivered in both approaches seem to indicate the correctness of the vision articulated above.

5. Geometric multiscale analysis 'with Calderón'

In harmonic analysis since the 1970's there have been a number of important applications of decompositions based on *parabolic dilations*

$$f_a(x_1,x_2) = f_1(a^{1/2}x_1, ax_2),$$

so called because they leave invariant the parabola $x_2 = x_1^2$. Calderón himself used such dilations [4] and exhibited a reproducing formula where the scale variable acted

through such dilations. Note that in the above equation the dilation is always twice as strong in one fixed direction as in the orthogonal one.

At the same time, decompositions began to be used based on **directional parabolic dilations** of the form

$$f_{a,\theta}(x_1, x_2) = f_a(R_\theta(x_1, x_2)').$$

Such dilations (essentially) leave invariant curves defined by quadratic forms with θ as one of the principal directions. For example, Charles Fefferman in effect used decompositions based on parabolic scaling in his study of Bochner-Riesz summability citeFefferman. Elias Stein used decompositions exhibiting parabolic scaling in studying oscillatory integrals in the 1970's and 1980's [45]. In the 1990's, Jean Bourgain, Hart Smith, Chris Sogge, and Elias Stein found applications in the study of oscillatory integrals and Fourier Integral operators.

The principle of parabolic scaling leads to a meaningful decomposition reminiscent of the continuous wavelet transform, only with a much more strongly directional character. This point has been developed in a recent article of Hart Smith [37], who defined a continuous wavelet transform based on parabolic scaling, a notion of directional molecule, showed that FIO's map directional molecules into directional molecules, and showed that FIO's have a sparse representation in a discrete decomposition. For this expository work, we have developed what seems a conceptually simple, perhaps novel way of approaching this topic, which we hope will be accesible to non-experts. Details underlying the exposition are available from [26].

5.1. Continuous directional multiscale analysis

We will work exclusively in $\mathbf{R}^2$, although everything generalizes to higher dimensions. Consider a family of directional wavelets with three parameters: scale $a > 0$, location $b \in \mathbf{R}^2$ and orientation $\theta \in [0, 2\pi)$. The orientation and location parameters are defined by the obvious rigid motion

$$\psi_{a,b,\theta} = \psi_{a,0,0}(R_\theta(x - b))$$

with R_θ the 2-by-2 rotation matrix effecting planar rotation by θ radians. At fine scales, the scale parameter a acts in a slightly nonstandard fashion based on parabolic dilation, *in the polar Fourier domain.* We pick a wavelet $\psi_{1,0}$ with $\hat{\psi}$ of compact support away from 0, and a bump $\phi_{1,0}$ supported in $[-1, 1]$. Here $\psi_{1,0}$ should obey the usual admissibility condition and $\|\phi\|_2 = 1$. At sufficiently fine scales (say $a < 1/2$) we define the directional wavelet by going to polar coordinates (r, ω) and setting

$$\hat{\psi}_{a,0,0}(r, \omega) = \hat{\psi}_{a,0}(r) \cdot \phi_{a^{1/2},0}(\omega), \quad a < a_0.$$

In effect, the scaling is parabolic in the polar variables r and ω, with ω being the 'thin' variable; thus in particular the wavelet $\psi_{a,0,0}$ is not obtainable by affine change-of-vartiables on $\psi_{a',0,0}$ for $a' \neq a$. We omit description of the transform at coarse scales, and so again ignore low frequency adjustment terms. Note that it is

correct to call these wavelets directional, since they become increasingly needle-like at fine scales.

Equipped with such a family of high-frequency wavelets, we can define a *Directional Wavelet Transform*

$$DW(a,b,\theta) = \langle \psi_{a,b,\theta}, f \rangle, \quad a > 0, b \in \mathbf{R}^2, \theta \in [0, 2\pi)$$

It is easy to see that we have a Calderón-like reproducing formula, valid for high-frequency functions:

$$f(x) = \int DW(a,b,\theta)\psi_{a,b,\theta}(x)\mu(da\, db\, d\theta)$$

and a Parseval formula for high-frequency functions:

$$\|f\|_{L^2}^2 = \int DW(a,b,\theta)^2 \mu(da\, db\, d\theta)$$

in both cases, μ denotes the reference measure $\frac{db}{a^{3/2}} \frac{d\theta}{a^{1/2}} \frac{da}{a}$.

Based on this transform, we can define seminorms reminiscent of Besov and Triebel seminorms in wavelet analysis; while it is probably a major task to prove that thse give well-founded spaces, and such work has not yet been done (for the most part), it still seems useful to use these as a tool measuring the distribution of a function's 'content' across scale, location and direction. We get a directional **Besov**-analog $DB^{\sigma}_{p,q}$: integrating over locations and orientations first

$$\left(\int \left(\int (|DW(a,b,\theta)|a^{-s})^p \frac{d\theta}{a^{1/2}} \frac{db}{a^{3/2}} \right)^{q/p} \frac{da}{a^2} \right)^{1/p}$$

and a **Triebel**-analog $DF^{\sigma}_{p,q}$ by integrating over scales first

$$\left(\int \left(\int (|DW(a,b,\theta)|a^{-s})^q \frac{da}{a^{1+2q/p}} \right)^{p/q} d\theta db \right)^{1/p}.$$

In both cases we take $s = \sigma - 3/2(1/p - 1/2)$. (There is the possibility of defining spaces using a third index (eg $B^{\sigma}_{p,q,r}$) corresponding to the L^r norm in the θ variable, but we ignore this here). As usual, the above formulas can only provide norms for high-frequency functions, and would have to be modified at coarse scales if any low frequencies were present in f. As in the case of the continuous wavelet transform for $\mathbf{R}$, there is some heuristic value in considering the transform as measuring finite directional differences e.g. $f(b+ae_\theta) - 2f(b) + f(b-ae_\theta)$, where $e_\theta = (\cos(\theta), \sin(\theta))'$; however this view is ultimately misleading. It is better to think of the transform as comparing the difference between polynomial approximation localized to two different rectangles, one of size a by $\sqrt{a}$ and the other, concentric and co-oriented, of size $2a$ by $\sqrt{2a}$.

The transform is actually performing a kind of microlocal analysis of f far more subtle than what is possible by simple difference/differential expressions. Indeed,

consider the Heaviside $H(x) = 1_{\{x_1>0\}}$; then at fine scales $DW(a,0,\theta) = 0$ for $|\theta| > \sqrt{a}$ and $DW(a,0,0) \approx a^{3/4}$ for $|\theta| \ll \sqrt{a}$, so that $\int_0^{2\pi} |DW(a,0,\theta)|d\theta \leq Ca^{5/4}$ as $a \to 0$. In short, DW is giving very precisely the orientation of the singularity. Moreover, for $b \neq (0,x_2)'$, $\int_0^{2\pi} |DW(a,0,\theta)|d\theta \to 0$ rapidly as $a \to 0$. So the transform is localizing the singularity quite well at fine scales, in a way that is difficult to imagine simple differences being able to do. Interpreting the above observations, we learn that a smoothly windowed Heaviside $f(x) = H(x)e^{-x^2}$ belongs in $DB^0_{\infty,\infty}$ but not in any better space $DB^\sigma_{\infty,\infty}$, $\sigma > 0$, while it belongs in $DB^1_{1,\infty}$ and not in any better space $DB^\sigma_{1,\infty}$, $\sigma > 1$. The difference between the critical indices in these cases is indicative of the sensitivity of the $p = 1$ seminorms to sparsity. Continuing in this vein, we have that for weak ℓ^p embeddings, for each $\eta > 0$

$$\mu\{(a,b,\theta) : |DW(a,b,\theta)| > \epsilon\} \leq C\epsilon^{-(3/2-\eta)}$$

so that the space-scale-direction plane for the (windowed) Heaviside is almost in $L^{2/3}(\mu) \sim DB^{3/2}_{2/3,2/3}$; the Heaviside has something like 3/2-derivatives. In comparison, the wavelet expansion of the Heaviside is only in ℓ^1, so the expansion is denser and 'more irregular' from the wavelet viewpoint than from the directional wavelet viewpoint. For comparison, the Dirac mass δ belongs at best to $B^{-1}_{\infty,\infty}$ and $B^0_{1,\infty}$ while it belongs at best to $DB^{-3/2}_{\infty,\infty}$ and $DB^{-1/2}_{1,\infty}$. The 'point singularity' is more regular from the wavelet viewpoint than from the directional wavelet viewpoint, while the Heaviside is more regular from the directional wavelet viewpoint than from the wavelet viewpoint. In effect, the Dirac 'misbehaves in every direction', while the Heaviside misbehaves only in one direction, and this makes a big difference for the directional wavelet transform.

There are two obvious special equivalences: first, $L^2 \sim DF^0_{2,2} \sim DB^0_{2,2}$ and L^2 Sobolev $W^m_2 \sim DF^m_{2,2} \sim DB^m_{2,2}$. There are in general no other L^p equivalences. Outside the L^2 Sobolev scale, the only equivalence with a previously proposed space is with Hart Smith's "Hardy Space for Fourier Integral Operators" [37]: $\mathcal{H}^1_{FIO} \sim DF^0_{1,2}$. This space has a molecular decomposition into *directional molecules*, which are functions that, at high frequency, are roughly localized in space to an a by $\sqrt{a}$ rectangle and roughly localized in frequency to the dual rectangle rotated 90 degrees, using traditional ways of measuring localization, such as boundedness of moments of all orders in the two principal directions. Under this qualitative definition of molecule, Smith showed that $\mathcal{H}^1_{FIO}$ has a molecular decomposition $f = \sum_Q A_Q m_Q(x)$ in which the coefficients obey an ℓ^1 norm summability condition $\sum |A_Q| 2^{j3/4} \leq 1$ when the directional molecules are L^2 normalized. This is obviously the harbinger for a whole theory of directional molecular decompositions.

More generally, one can make a molecular decomposition of the directional Besov and directional Triebel classes by discretizing the directional wavelet transform according to tiles $Q = Q(j,k_1,k_2,\ell)$ which obey the following desiderata:

- In tile $Q(j,k_1,k_2,\ell)$, scale a runs through a dyadic interval $2^{-j} > a \geq 2^{-(j+1)}$.
- At scale 2^{-j}, locations run through rectangularly shaped regions with aspect ratio roughly 2^{-j} by $2^{-j/2}$.

- The location regions are rotated consistent with the orientation $b \approx R_{\theta_\ell}(k_1/2^j, k_2/2^{j/2})$.
- The tile contains orientations running through $2\pi\ell/2^{j/2} \le \theta < 2\pi(\ell+1)/2^{j/2}$.

Note again that for such tiles $\mu(Q) \approx 1$. Over such tiles different values of $DW(a, b\theta)$ are roughly comparable and different wavelets $\psi_{a,b,\theta}$ as well. Hence it is sensible to decompose

$$\begin{aligned} f(x) &= \int DW(a,b,\theta)\psi_{a,b,\theta}(x)\mu(dadbd\theta) \\ &= \sum_Q \int_Q DW(a,b,\theta)\psi_{a,b,\theta}(x)\mu(dadbd\theta) \\ &= \sum_Q M_Q(x), \qquad M_Q(x) = \int_Q DW(a,b,\theta)\psi_{a,b,\theta}(x)\mu(dadbd\theta) \\ &= \sum_Q A_Q m_Q(x), \qquad A_Q = \|DW(a,b,\theta)\|_{L^2(Q)} \end{aligned}$$

Morever, for any decomposition into directional molecules (not just the approach above), the appropriate sequence norm of the amplitude coefficients gives control of the corresponding directional Besov or directional Triebel norm. It is then relatively immediate that one can define sequence space norms for which we have the norm equivalences

$$\|f\|_{DB^\sigma_{p,q}} \asymp \|(A_Q)_Q\|_{db^\sigma_{p,q}}, \qquad \|f\|_{DF^\sigma_{p,q}} \asymp \|(A_Q)_Q\|_{df^\sigma_{p,q}} \tag{5.5}$$

where we again omit discussion of low frequency terms. The sequence space equivalence $db^0_{2,2} \sim df^0_{2,2} \sim \ell^2$ are trivial. An interesting equivalence of relevance to the Heaviside example above is $db^{3/2}_{2/3,2/3} \sim \ell^{2/3}$, so that, again, a smoothness space with "$p < 1$" is equivalent to an ℓ^τ ball with $\tau < 1$.

Hart Smith made the crucial observation that the molecules for the Smith space are invariant under diffeomorphisms. That is, if we take a C^∞ diffeomorphism ϕ, and a family of $\mathcal{H}^1_{FIO}$ molecules (such as $m_Q(x)$), then every $\tilde{m}_Q(x) = m_Q(\phi(x))$ is again a molecule, and the sizes of moments defining the molecule property are comparable for m_Q and for $\tilde{m}_Q$. It follows that $\mathcal{H}^1_{FIO}$ is invariant under diffeomorphisms of the base space. His basic lemma underlying this proof was strong enough to apply to invariance of directional molecules in every one of the directional Besov and directional Triebel classes. Hence directional Besov and directional Triebel classes are invariant under diffeomorphisms of the base space.

This invariance enables a very simple calculation, suggesting that the *directional wavelet transform sparsifies objects with singularities along smooth curves*, or at least sparsifies such objects to a greater extent that does the ordinary wavelet transform. Suppose we analyse a function f which is smooth away from a discontinuity along a straight line; then the Heaviside calculation we did earlier shows that most directional wavelet coefficients are almost in weak $L^{2/3}$. Now since objects with linear singularities have $\ell^{2/3+\epsilon}$ boundedness of amplitudes in a molecular decomposition, and directional molecules are diffeomorphism invariant, this sparsity

condition is invariant under diffeomorphisms of the underlying space. It follows that an object which is smooth away from a discontinuity along a smooth curve should also have molecular amplitudes in $\ell^{2/3+\epsilon}$.

This sparsity argument suggests that directional wavelets outperform wavelets for representing such geometric objects. Indeed, for $\eta > 0$ there is an $\epsilon > 0$ so that $\ell^{2/3+\epsilon}$ boundedness of directional wavelet molecular amplitudes shows that approximation by sums of N directional molecules allows a squared-L^2 approximation error of order $O(N^{-2+\eta})$, whereas wavelet coefficients of such objects are only in ℓ^1, so sums of N wavelets only allow squared-L^2 approximation error of size $O(N^{-1})$.

5.2. Stylized applications

The above calculations about sparsification of objects with curvilinear singularities suggests the possibility of using the directional wavelet transform based on parabolic scaling to pursue counterparts of all the various classical wavelet applications mentioned in Section 3: nonlinear approximation, data compression, noise removal, and fast computations. It further suggests that such directional wavelet methods might outperform calssical wavelets – at least for objects containing singularities along smooth curves, i.e. edges.

5.2.1. First discretization: curvelets

To develop applications, molecular decomposition is (once again) not enough: some sort of rigid decomposition needs to be developed; an orthobasis, for example.

Candès and Donoho [7] developed a tight frame of elements exhibiting parabolic dilations which they called *curvelets*, and used it to systematically develop some of these applications. A side benefit of their work is knowledge that the transform is essentially optimal, i.e. that there is no fundamentally better scheme of nonlinear approximation. The curvelet system has a countable collection of generating elements $\gamma_\mu(x_1, x_2)$, $\mu \sim (a_j, b_{k_1,k_2}, \theta_\ell, t_m)$ which code for scale, location, and direction. They obey the usual rules for a tight frame, namely, the reconstruction formula and the Parseval relation:

$$f = \sum_\mu \langle \gamma_\mu, f \rangle \gamma_\mu, \qquad \|f\|_2^2 = \sum_\mu \langle \gamma_\mu, f \rangle^2.$$

The transform is based on a series of space/frequency localizations, as follows.

- Bandpass filtering. The object is separated out into different dyadic scale subbands, using traditional bandpass filtering with passband centered around $|\xi| \in [2^j, 2^{j+1}]$.
- Spatial localization. Each bandpass object is then smoothly partitioned spatially into boxes of side $2^{-j/2}$.
- Angular localization. Each box is analysed by ridgelet transform.

The frame elements are essentially localized into boxes of side 2^{-j} by $2^{-j/2}$ at a range of scales, locations, and orientations, so that it is completely consistent with the molecular decomposition of the directional Besov or directional Fourier classes.

However, unlike the molecular decomposition, the coefficients are linear in f and the frame elements are fixed elements. Moreover, an algorithm for application to real data on a grid is relatively immediate.

5.2.2. Nonlinear approximation

In dimension 2, the analog to what was called free knot spline approximation is approximation by piecewise polynomials on triangulations with N pieces. This idea has generated a lot of interest but frustratingly few hard results. For one thing, it is not obvious how to build such triangulations in a way that will fulfill their apparent promise, and in which the resulting algorithm is practical and possible to analyze.

Here is a class of two-dimensional functions where this scheme might be very attractive. Consider a class $\mathcal{F}$ of model 'images' which exhibit discontinuities across C^2 smooth curves. These 'images' are supposed to be C^2 away from discontinuity. Moreover, we assume uniform control both of the C^2 norm for the discontinuity curve and smooth function. One can imagine that very fine needle-like triangles near curved discontinuities would be valuable; and this is indeed so, as [27] shows; in an ideal triangulation one geta a squared error converging at rate N^{-2} whereas adaptive quadtrees and other simpler partitioning schemes give only N^{-1} convergence. Moreover, this rate is optimal, as shown in [27], if we allow piecewise smooth approximation on essentially arbitrary triangulations with N pieces, even those designed by some as yet unknown very clever and very nonlinear algorithm, we cannot in general converge to such objects faster than rate N^{-2}.

Surprisingly, a very concrete algorithm does almost this well: simply thresholding the curvelet coefficients. Candès and Donoho have shown the following [8]

Theorem: *The decreasing rearrangement of the frame coefficients in the curvelet system obeys the following inequality for all $f \in \mathcal{F}$:*

$$|\alpha|_{(k)} \le Ck^{-3/2}\log^{3/2}(k), \qquad k \ge 1.$$

This has exactly the implication one would have hoped for from the molecular decomposition of directional Besov classes: the frame coefficients are in $\ell^{2/3+\epsilon}$ for each $\epsilon > 0$. Hence, we can build an approximation to a smooth object with curvilinear discontinuity from N curvelets with squared L^2-error $\log^3(N) \cdot N^{-2}$; as mentioned earlier, Wavelets would give squared L^2-error $\ge cN^{-1}$.

In words: approximation by sums of the N-biggest curvelet terms does essentially as well in approximating objects in $\mathcal{F}$ as free-triangulation into N regions. In a sense, the result is analogous to the result mentioned above in Section 3.2.1 comparing wavelet thresholding to nonlinear spline approximation, where we saw that approximation by the N-biggest amplitude wavelet terms does as well as free-knot splines with N knots. There has been a certain amount of talk about the problem of characterizing approximation spaces for approximation by N arbitrary triangles; while this problem seems very intractable, it is clear that the directional Besov classes provide what is, at the moment, the next best thing.

5.2.3. Data compression

Applying just the arguments already given in the wavelet case show that the result of L^2 nonlinear approximation by curvelets, combined with simple quantization, gives near-optimal compression of functions in the class $\mathcal{F}$ above, i.e. the number of bits in the compressed representation is optimal to within some polylog factor. This seems to promise some interesting practical coders someday.

5.2.4. Noise removal

The results on nonlinear approximation by thresholding of the curvelet coefficients have corresponding implications in statistical estimation. Suppose that we have noisy data according to the white noise model

$$Y(dx_1, dx_2) = f(x_1, x_2)dx_1dx_2 + \epsilon W(dx_1, dx_2)$$

where W is a Wiener sheet. Here f comes from the same 'Image Model' $\mathcal{F}$ discussed earlier, of smooth objects with discontinuities across C^2 smooth curves. We measure risk by Mean Squared Error, and consider the estimator that thresholds the curvelet coefficients at an appropriate (roughly $2\sqrt{\log(\epsilon^{-1})}$) multiple of the noise level. Emmanuel Candès and I showed the following [9]:

Theorem: *Appropriate thresholding of curvelet coefficients gives nearly the optimal rate of convergence; with polylog*(ϵ) *a polynomial in* $\log(1/\epsilon)$*, the estimator* $\hat{f}^{CT}$ *obeys*

$$R_\epsilon(f, \hat{f}^{CT}) \le \mathit{polylog}(\epsilon) \cdot \min_{\hat{f}} \max_{f \in \mathcal{F}} R_\epsilon(f, \hat{f}).$$

Hence, in this situation, curvelet thresholding outperforms wavelet thresholding at the level of rates: $O(\mathit{polylog}(\epsilon) \cdot \epsilon^{4/3})$ vs $O(\epsilon)$. Similar results can be developed for other estimation problems, such as the problem of Radon inversion. There the rate comparison is $\mathit{polylog}(\epsilon) \cdot \epsilon^{4/5}$ vs $\log(1/\epsilon) \cdot \epsilon^{2/3}$; [9]. In empirical work [44, 10], we have seen visually persuasive results.

5.2.5. Improved discretization: directional framelets

The curvelet representation described earlier is a somewhat awkward way of obtaining parabolic scaling, and also only indirectly related to the continuum directional wavelet transform. Candès and Guo [10] suggested a different tight frame expansion based on parabolic scaling. Although this was not introduced in such a fashion, for this exposition, we propose an alternate way to understand their frame, simply as discretizing the directional wavelet transform in a way reminiscent of (3.1); for details, see [26]. Assuming a very specific choice of directional wavelet, one can get (the fine scale) frame coefficients simply by sampling the directional wavelet transform, obtaining a decomposition

$$f = \sum_{j,k,l} DW(2^{-j}, b^{j,\ell}_{k_1,k_2}, 2\pi l/2^{j/2})\psi_{2^{-j},k/2^j,2\pi\ell/2^{j/2}} = \sum_{j,k,l} \alpha_{j,k,l}\psi_{j,k,l}, \text{say ;}$$

(as usual, this is valid as written only for high-frequency functions). In fact this can yield a tight frame, in particular the Parseval relation $\sum_{j,k,l} \alpha^2_{j,k,l} = \|f\|^2_{L^2}$.

This has conceptual advantages: a better relationship to the continuous directional wavelet transform and perhaps an easier path to digital representation. In comparison with the original curvelets scheme, curvelets most naturally organizes matters so that 'within' each location we see all directional behavior represented, whereas directional framelets most naturally organize matters so that 'within' each orientation we see all locations represented.

5.2.6. Operator representation

Hart Smith, at the Berlin ICM, mentioned that decompositions based on parabolic scaling were valuable for understanding Fourier Integral Operators (FIO's) [38]; in the notation of our paper, his claim was essentially that FIO's of order zero operate on fine-scale directional molecules approximately by performing well-behaved affine motions – roughly, displacement, scaling and change of orientation. Underlying his argument was the study of families of elements generated from a single wavelet by true affine parabolic scaling $\phi_{a,b,\theta}(x) = \phi(P_a \circ R_\theta \circ S_b x)$ where $P_a = diag(a, \sqrt{a})$ is the parabolic scaling operator and $S_b x = x - b$ is the shift. Smith showed that if T is an FIO of order 0 and ϕ is directionally localized, the kernel

$$K_{a,b,\theta}^{a',b',\theta'} = \langle \phi_{a,b,\theta}, T\phi_{a',b',\theta'} \rangle$$

is rapidly decaying in its entries as one moves away from 'the diagonal' in an appropriate sense.

Making this principle more adapted to discrete frame representations seems an important priority. Candès and Demanet have recently announced [11] that actually, *the matrix representation of FIOs of order* 0 *in the directional framelet decomposition is sparse.* That is, each row and column of the matrix will be in ℓ^p for each $p > 0$. in a directional wavelet frame. This observation is analogous in some ways to Meyer's observation that the orthogonal wavelet transform gives a sparse representation for Calderón-Zygmund operators. Candès has hopes that this sparsity may form some day the basis for fast algorithms for hyperbolic PDE's and other FIO's.

5.3. Applications

The formalization of the directional wavelet transform and curvelet transform are simply too recent to have had any substantial applications of the 'in daily use by thousands' category. Serious deployment into applications in data compression or statistical estimation is still off in the future.

However, the article [29] points to the possibility of immediate effects on research activity in computational neuroscience, simply by generating new research hypothesis. In effect, if vision scientists can be induced to consider these new types of image representation, this will stimulate meaningful new experiments, and reanalyses of existing experiments.

To begin with, for decades, vision scientists have been influenced by mathematical ideas in framing research hypotheses about the functioning of the visual cortex, particular the functioning of the V1 region. In the 1970's, several authors

suggested that the early visual system does Fourier Analysis; by the 1980's the cutting edge hypothesis became the suggestion that the early visual system does Gabor Analysis; and by the 1990's, one saw claims that the early visual system does a form of wavelet analysis. While the hypotheses have changed over time, the invariant is that vision scientists have relied on mathematics to provide language & intellectual framework for their investigations. But it seems likely that the hypotheses of these previous decades are incomplete, and that to these should be added the hypothesis that the early visual system performs a directional wavelet transform based on parabolic scaling. During my Plenary Lecture, biological evidence was presented consistent with this hypothesis, and a proposal was made that future experiments in intrinsic optical imaging of the visual cortex ought to attempt to test this hypothesis. See also [29].

6. Geometric multiscale analysis 'without Calderón'

In the last section we considered a kind of geometric multiscale analysis employing a Calderón-like formula. In the ICM Lecture we also considered dispensing with the need for Calderón formulas, using a cruder set of multiscale tools, but one which allows for a wide range of interesting applications – very different from the applications based on analysis/synthesis and Parseval. Our model for how to get started in this direction was Peter Jones' travelling salesman problem. Jones considered instead a countable number of points $X = \{x_i\}$ in $[0,1]^2$ and asked: when can the points of X be connected by a finite length (rectifiable) curve ? And, if they can be, what is the shortest possible length? Jones showed that one should consider, for each dyadic square Q such that the dilate $3Q$ intersects X, the width w_Q of the thinnest strip in the plane containing all the points in $X \cap 3Q$, and define $\beta_Q = w_Q/diam(Q)$ the proportional width of that strip, relative to the sidelength of Q. As $\beta_Q = 0$ when the data lie on a straight line, this is precisely a measure of how close to linear the data are over the square Q. He proved the there is a finite-length curve Γ visiting all the points in $X = \{x_i\}$ iff $\sum \beta_Q^2 diam(Q) < \infty$. I find it very impressive that analysis of the number of points in strips of various widths can reveal the existence of a rectifiable curve connecting those points. In our lecture, we discussed this idea of counting points in anistropic strips and several applications in signal detection and pattern recognition [1, 2], with applications in characterizing galaxy clustering [34]. We also referred to interesting work such as Gilad Lerman's thesis [40], under the direction of Coifman and Jones, and to [30], which surveys a wide range of related work. Look to [26] for an extended version of this article covering such topics.

7. Conclusion

Important developments in 'pure' harmonic analysis, like the use of parabolic scaling for study of convolution operators and FIOs, or the use of anisotropic strips for analysis of rectifiable measures, did not arise because of applications to our

developing 'information society', yet they seem to have important stylized applications which point clearly in that direction. A number of enthusiastic applied mathematicians, statisticians, and scientists are attempting to develop true 'real world' applications.

At the same time, the fruitful directions for new kinds of geometric multiscale analysis and the possible limitations to be surmounted remain to be determined. Stay tuned!

8. Acknowledgements

The author would like to thank Emmanuel Candès, Raphy Coifman, Peter Jones, and Yves Meyer for very insightful and inspiring discussions. This work has been partially supported by National Science Foundation grants DMS 00-77261, 98–72890 (KDI), and DMS 95–05151.

References

[1] Arias, E., Donoho, D.L., Huo, X. and Tovey, C. (2002) 'Connect-the-Dots': How many random points can touch a smooth curve or surface? Manuscript.

[2] Arias, E., Donoho, D.L., and Huo, X. (2002) Multiscale Detection of Geometric Objects Buried in Noisy Images. Manuscript.

[3] Beylkin G., Coifman R., Rokhlin V. (1991) Fast Wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.*, 44, 141–183.

[4] Calderón, Alberto. P. (1977) An atomic decomposition of distributions in parabolic H^p spaces. *Advances in Math.* **25** , no. 3, 216–225.

[5] Candés, Emmanuel J. (1998) *Ridgelets: Theory and Applications.* Ph.D. Thesis, Department of Statistics, Stanford University.

[6] Candès, Emmanuel J. Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.* 6 (1999), no. 2, 197–218.

[7] Candès, E.J. and Donoho, D.L. (2000) Curvelets: a surprisingly effective nonadaptive representation of objects with edges. in *Curve and Surface Fitting: Saint-Malo 1999* Albert Cohen, Christophe Rabut, and Larry L. Schumaker (eds.) Vanderbilt University Press, Nashville, TN.

[8] Candès, E.J. and Donoho, D.L. (2000) Curvelets: sparse representation of objects with edges via parabolic scaling. Manuscript.

[9] Emmanuel J. Candès and David L. Donoho, Recovering edges in ill-posed inverse problems: optimality of curvelet frames. *Ann. Statist.* **30** (2002), no. 3, 784–842.

[10] E. J. Candès and F. Guo (2002). New Multiscale Transforms, Minimum Total Variation Synthesis: Applications to Edge-Preserving Image Reconstruction, to appear in *Signal Processing.*

[11] E.J. Candès and L. Demanet (2002) Presentation at Foundations of Computational Mathematics Workshop 2002.

[12] A.Cohen, I.Daubechies, M.Orchard and O.Gulieriz, "On the importance of

combining wavelet-based non-linear approximation with coding strategies", to appear in *IEEE Trans. Inf. Theory*, 2002.
[13] R.R. Coifman (1974) An Atomic Decomposition of H^p. *Studia Math.* 51.
[14] R. Coifman and Guido Weiss. (1977) Extensions of Hardy spaces and their use in analysis. **Bull. Amer. Math. Soc. 83**, no. 4, 569–645.
[15] Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.*, **41**, 909–996.
[16] Daubechies, I. (1992) *Ten Lectures on Wavelets.* CBMS-NSF Series in Applied Mathematics, No. 61. SIAM Philadelphia.
[17] G. David and S. Semmes (1993) *Analysis of and on Uniformly Rectifiable Sets.* Math Surveys and Monographs **38**, Providence: AMS.
[18] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Am. Math. Soc.* **305** (1), 397-414.
[19] RA DeVore, B. Jawerth, BJ Lucier (1992), Image compression through wavelet transform coding, *IEEE Trans. Info. Theory*, **38**, No. 2, 719–746, March 1992.
[20] Donoho, D.L. (1993) Unconditional Bases are Optimal Bases for Data Compression and for Statistical Estimation. *Applied and Computational Harmonic Analysis*, **1**, Dec. 1993, pp. 100-115.
[21] Donoho, D.L. (1995) Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis.* **2**, May 1995, pp. 101-126.
[22] D.L. Donoho (1996) Unconditional Bases and Bit-Level Compression. *Applied and Computational Harmonic Analysis* **3** 388-392.
[23] Donoho, D.L. (1999) Tight Frames of k-Plane Ridgelets and the Problem of Representing d-dimensional singularities in $\mathbf{R}^n$. *Proc. Nat. Acad. Sci. USA*, **96**, 1828-1833.
[24] Donoho, D.L. (2000) Orthonormal Ridgelets and Linear Singularities. *SIAM J. Math Anal.* Vol. 31 Number 5 pp. 1062-1099.
[25] D.L. Donoho (2000) Counting Bits with Kolmogorov and Shannon. Manuscript.
[26] *http://www-stat.stanford.edu/~donoho/Lectures/ICM2002.*
[27] Donoho, D.L. (2001) Sparse Components of Images and Optimal Atomic Decomposition. *Constructive Approximation*, **17** , no. 3, 353–382.
[28] Donoho, D.L., and Duncan, M.R. (2000) Digital Curvelet Transform: Strategy, Implementation, Experiments. in *Wavelet Applications VII*, H.H. Szu, M. Vetterli, W. Campbell, and J.R. Buss, eds. (Proc. Aerosense 2000, Orlando, Fla.), SPIE vol 4056, pp. 12-29. SPIE: Bellingham Washington, 2000.
[29] Donoho, D.L. and Flesia, A.G. (2001) Can Recent Advances in Harmonic Analysis Explain Recent Findings in Natural Scene Statistics? *Network: Computation in Neural Systems*, **12** (no.3), Aug. 2001. 371–393.
[30] Donoho, D.L. and Huo, Xiaoming. (2001) Beamlets and Multiscale Image Analysis. in *Multiscale and Multiresolution Methods,* T. J. Barth and T. F. Chan and R. Haimes eds. Lecture Notes in Computational Science and Engineering **20**, Springer-Verlag, 149–196.
[31] Donoho, D.L., Johnstone I.M., Kerkyacharian, G., and Picard, D. (1995) Wavelet Shrinkage: Asymptopia? *Journ. Roy Stat. Soc. Ser B*, **57**, no. 2,

1995, 301–369.
[32] Donoho, D.L. and Johnstone, I.M. (1998) Minimax estimation via wavelet shrinkage. *Ann.Statist.* **26**, 3, 879–921.
[33] Donoho, D.L., Vetterli, M., DeVore, R.A., and Daubechies, I. (1998) Data Compression and Harmonic Analysis. *IEEE Trans. Info. Thry.* **44**, 6, 2435-2476.
[34] D.L. Donoho, O. Levi, J.L. Starck, and V. Martinez. (2002) Multiscale Geometric Analysis of Galaxy Catalogs. to appear *Astronomical Telescopes and Instrumentation*, SPIE: Bellingham Washington 2002.
[35] Fefferman, Charles (1973). A note on spherical summation multipliers. *Israel J. Math.* **15**, 44–52.
[36] M. Frazier, B. Jawerth, and G. Weiss (1991) *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, **79**. American Math. Soc.: Providence, RI.
[37] Smith, Hart F. (1998) A Hardy space for Fourier integral operators. *J. Geom. Anal.* **8**, no. 4, 629–653.
[38] Smith, Hart F. (1998) Wave equations with low regularity coefficients, *Documenta Mathematica*, Extra Volume ICM 1998, II, 723-730.
[39] P. W. Jones (1990) "Rectifiable Sets and the Travelling Salesman Problem." *Inventiones Mathematicae*, **102**, 1–15.
[40] G. Lerman (2000) Geometric Transcriptions of Sets and Their Applications to Data Analysis. Ph.D. Thesis, Yale University, Department of Mathematics.
[41] Mallat, S. (1989b) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.
[42] Meyer, Y. (1990a) *Ondelettes.* Paris: Hermann. English translation published by Cambridge University Press.
[43] Meyer, Y. (1990b) *Operateurs de Calderón et Zygmund.* Paris: Hermann.
[44] J. L. Starck, E. J. Candès and D. L. Donoho (2000). The Curvelet Transform for Image Denoising. To appear *IEEE Transactions on Signal Processing*, **11**, 6, pp 670-684, 2002.
[45] Elias M. Stein (1993) *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals.* Princeton University Press, Princeton, NJ.
[46] G. Weiss, R. Laugesen, E. Wilson, and N. Weaver, A Generalized Caldern Reproducing Formula and its Associated Continuous Wavelets, appear *J. of Geom. Anal.*
[47] G. Weiss, P. Gressman, D. Labate, E. Wilson. Affine, Quasi-Affine and Co-Affine Wavelets, to appear *Beyond Wavelets*, J. Stoeckler and G. V. Welland (Eds), Academic Press.

ICM 2002 · Vol. I · 235–244

Knotted Solitons

L. D. Faddeev*

Abstract

The dynamical model on 3+1 dimensional space-time admitting soliton solutions is discussed. The proposal soliton is localized in the vicinity of a closed contour, which could be linked and/or knotted. The topological charge is Hopf invariant. Some applications in realistic physical systems are indicated.

2000 Mathematics Subject Classification: 35Q51, 35S35, 65C20, 81V25.
Keywords and Phrases: Soliton, Knot, Hopf invariant.

1. Introduction

The term "soliton" entered applied mathematics in 1965. It was coined by M. Kruskal and N. Zabusky for a special solution of nonlinear Korteweg-de Vries (KdV) equation, depicting solitary wave [1]. Use of convention of particle physics language shows that the author envisioned the particle-like interpretation for the object which they called soliton.

The attention of mathematical physicists to solitons was attracted after the inverse scattering method was devised by G. Gardner, J. Green, M. Kruskal and R. Miura for solving the KdV equation [2] and its extension to Nonlinear Schroedinger Equation was found by V. Zakharov and A. Shabat [3]. In the 1970's this method and its generalizations got a lot of attention and involved quite a few active participants. Rather complete review can be found in [4]. In the end of that decade the quantum variant of the method was constructed and particle-like interpretations of solitons got natural confirmation in terms of quantum field theory, see review in [5]. The mathematical structure of the quantum method was deciphered in pure algebraic way leading in the 1980's to notion of quantum groups with new applications in pure mathematics and mathematical physics.

The value of solitons for the particle physics consists in the possibility of going beyond the paradigm of the perturbation theory. Indeed, soliton solutions correspond to full nonlinear equations and disappear in their linearized form. Characteristic for solitons is that they interact strongly if the excitations of the linearized

*St. Petersburg Department of Steklov Mathematical Institute, Russian Academy of Sciences, Russia. E-mail: faddeev@pdmi.ras.ru

fields interact weakly. Another attractive feature is the appearance of elementary topological characteristics for solitons topological charges.

This was understood already in the middle of 1970's by several groups as I underlined in my lectures, when I was touring USA in 1975 (see e. g. [6]). However, all these tantalizing features of solitons had one very important drawback: the developed methods applied only in 1 + 1 dimensional space-time.

Naturally the search for 3 + 1 dimensional generalizations became eminent. General considerations showed that many features of 1 + 1 dimensional systems, such as complete integrability and existence of exact many-particle solutions could not be generalized to 3 + 1 dimensions. However, the mere existence of "one-particle" soliton solutions was not excluded. One particular example was introduced by Skyrme in a pioneer paper [7] long before the soliton rush. Another example was proposed by G. 't Hooft and A. Polyakov in 1975 [8]. In the following years their solutions got real applications in nuclear and high energy physics.

In both examples the solitons are "point-like", namely their deviation from the vacuum is concentrated around central point in space. Moreover they have spherical symmetry, allowing the separation of variables in the corresponding equations, reducing them to ODE, which one can treat on a usual PC.

In my lectures [6], already mentioned, I proposed one more possibility for 3 + 1 dimensional system, allowing solitons. The model, which superficially looks as a slight modification of Skyrme model, has quite distinct features. The center of the would-be soliton is not a point, but a closed contour, possibly linked or knotted. However my proposal remained unnoticed. The reason was evident: the maximal symmetry for such a soliton is axial, reducing 3-dimensional nonlinear PDE to 2-dimensional one. Existing computers were not able to treat such a problem. Thus my proposal was in slumber for 20 years until my colleague Antti Niemi became interested and agreed to sacrifice a year to learn computing and devising the programm. The preliminary results published in [9] attracted the attention of professionals in computational physics and now we have an ample evidence, confirming my proposal [10], [11].

The development which followed showed unexpected universality of my model. The variables, used in it, were shown to enter the list of degrees of freedom for several systems, having realistic physical applications [12], [13].

In this talk I shall describe all these developments in detail. First I shall introduce the model, then briefly discuss its numerical treatment and finish with the description of the applications.

2. The field configurations and Hopf invariant

The space time is 4-dimensional Minkowski space M with linear coordinates x^μ, $\mu = 0, 1, 2, 3$, x^0 being time and x^k, $k = 1, 2, 3$ space variables. The field $\vec{n}(x)$ is defined on M and has values on 2-dimensional sphere $\mathbb{S}^2$:

$$\vec{n} : M \to \mathbb{S}^2.$$

The boundary condition on spatial infinity is introduced

$$\vec{n}|_{r=\infty} = \vec{n}_0, \tag{2.1}$$

where $r = ((x^1)^2 + (x^2)^2 + (x^3)^2)^{1/2}$ and $\vec{n}_0$ is a fixed vector, e.g. corresponding to the north pole

$$\vec{n}_0 = (0, 0, 1).$$

We shall consider mostly the time independent configurations, corresponding to a soliton at rest. The boundary condition (2.1) effectively compactifies the space $\mathbb{R}^3$, turning it into sphere $\mathbb{S}^3$, thus the stationary configurations realize the map

$$\vec{n} : \mathbb{S}^3 \to \mathbb{S}^2, \tag{2.2}$$

which are known to be classified by Hopf invariant, sort of topological charge.

In general, the density of topological charge is the zero component J_0 of the current J_μ, which is conserved

$$\partial_\mu J_\mu = 0$$

independently of the equations of motion. Mathematically it is more natural to use the 3-form J dual to 1-form $J^* = J_\mu \, \mathrm{d}\, x^\mu$ and define the topological charge as an integral of J over space section

$$Q = \int_{\mathbb{R}^3} J.$$

In our case the 3-form J is constructed as follows. The pull-back of the volume 2-form on $\mathbb{S}^2$ via map (2.2) defines the closed 2-form on the space time

$$H = H_{\mu\nu} \, \mathrm{d}\, x^\mu \wedge \mathrm{d}\, x^\nu,$$

where antisymmetric tensor $H_{\mu\nu}$ is expressed via field configuration $\vec{n}(x)$ as follows

$$H_{\mu\nu} = (\partial_\mu \vec{n} \times \partial_\nu \vec{n}, \vec{n}). \tag{2.3}$$

Here I use usual notations of vector analysis in 3-space. In fact H is exact

$$H = \mathrm{d}\, C$$

and current 3-form is given by

$$J = \frac{1}{4\pi} H \wedge C.$$

In more detail, we have the relations

$$H_{ik} = \partial_i C_k - \partial_k C_i$$

and

$$Q = \frac{1}{4\pi} \int \varepsilon_{ikj} H_{ik} C_j d^3 x.$$

For regular configurations Q gets integer values. This integer has a nice interpretation in the description of which I shall use the terminology of magnetostatic.

Tensor H_{ik} can be interpreted as a field strength of the stationary magnetic field in Maxwell theory. The corresponding lines of force are defined via equations

$$\frac{d}{ds}x_i = \frac{1}{2}\varepsilon_{ikj}H_{kj},$$

where s is a local parameter along the line. It is easy to see that components of $\vec{n}(x)$ along these lines are constant

$$\frac{d}{ds}\vec{n}(x) = 0,$$

giving two "integrals of the motion". In other words, the Maxwell lines of force are the preimages of points on $\mathbb{S}^2$ under the map (2.2). Hopf invariant is the intersection number of any pair of such lines.

All these facts are well known and can be found in textbooks (see e.g. [14]). However I decided to include them into my text to make it more selfcontained.

3. The dynamical model

I introduce the dynamical model by giving the relativistic action functional

$$\mathcal{A} = a\int(\partial_\mu\vec{n})^2 d^4x + b\int(H_{\mu\nu})^2 d^4x.$$

In the usual convention of high-energy physics $\mathcal{A}$ is dimensionless, so the parameter a has dimension $[\text{length}]^{-2}$ and parameter b is dimensionless. Corresponding static energy E has the same form as $\mathcal{A}$ with space-time coordinates substituted by space variables only

$$E = a\int(\partial_k\vec{n})^2 d^3x + b\int(H_{ik})^2 d^3x. \tag{3.4}$$

and has proper dimension $[\text{length}]^{-1}$. The structure of E is similar to that of Skyrme model, where the field variable having values in $\mathbb{S}^3$ is used and corresponding topological charge is just a degree of map.

Usual check based on the scale transformation is favorable for (3.4) in the same way as in Skyrme model. Indeed

$$E = E_2 + E_4,$$

where E_2 and E_4 are quadratic and quartic in derivatives of $\vec{n}$ correspondingly. Thus under scaling $x \to \lambda x$ we have

$$E_2 \to \lambda E_2\,, \quad E_4 \to \frac{1}{\lambda}E_4$$

and the virial theorem states that on the minimal configuration (if any)

$$E_2 = E_4.$$

In terms of quantum theory E_2 has a standard interpretation of the energy of nonlinear sigma-model whereas E_4 is rather exotic. On the contrary in the magnetic interpretation, mentioned above, E_4 is a natural term — it is just the Maxwell magnetic energy, whereas the nature of E_2 is not that clear. However in what follows the presence of both E_2 and E_4 is crucial for the existence of solitons as the scaling argument already showed.

This is confirmed also by a beautiful estimate, obtained in [15]

$$E \geq c|Q|^{3/4},$$

which shows that in the sectors with nonzero Q the minimum of energy is strongly positive. Thus the soliton solutions should be obtained by the minimizing of E with $Q \neq 0$ fixed.

Unfortunately until now there exists no proof of the compactness of the minimizing sequence in general case. For the case of axial symmetry uncouraging result are obtained in [16]. So the main argument for the evidence of solitons in my model is based on the numerical work.

4. Numerical work

To find the numerical evidence of the existence of localized solitons it is not necessary to solve the nonlinear elliptic equation, obtained by the variational principle

$$\frac{\delta E}{\delta \vec{n}} = 0. \tag{4.5}$$

Instead one can introduce an auxiliary time s and consider the parabolic equation

$$\frac{d\vec{n}}{ds} = \frac{\delta E}{\delta \vec{n}} \tag{4.6}$$

with initial value $\vec{n}_{\text{init}}$

$$\vec{n}|_{s=0} = \vec{n}_{\text{init}}$$

being a configuration with the prescribed Hopf invariant. Of course to simulate (4.6) on the computer one is to use some difference scheme. If for large s solution of (4.6) stabilizes it gives the solution of (4.5). In other words the soliton appears as an attractor for the evolution equation.

There are of course many important practical details how to discretize equation, how to take into account the normalization condition $\vec{n}^2 = 1$ and how to choose the initial configuration $\vec{n}_{\text{init}}$. The main papers [10] and [11] use different prescription for all this, however quite satisfactorily the final results coincide. I refer to these papers for the details of calculations and proceed to describe the results.

The iterative process was performed for the configuration with $Q = 1, 2, \ldots 7$. The results are as follows: for $Q = 1$ and $Q = 2$ the solutions are axial symmetric. The center line — the preimage of the point $n = (0, 0, -1)$ — is a circle. The surfaces $n_3 = \alpha$, $-1 < \alpha < 1$ are toroidal and they are spanned by the lines of force

wrapping the torus once for $Q = 1$ and twice for $Q = 2$. In other words the soliton can be viewed as a filament of lines of force, closed and twisted once or twice.

The solution for $Q = 3$ is similar but not axial symmetric any more, the corresponding "cable" is warped. For $Q = 4$ the soliton is a link of two twisted filaments. Especially beautiful case is $Q = 7$, the central line of the corresponding soliton is a trefoil knot.

The file [17] contains impressive moving pictures illustrating the convergence of the iterations. I plan to show these movies in my talk, but unfortunately can not do it in a written text.

Thus the numerical work gives the compelling evidence of the existence of string-like solitons in my model. There remains an important mathematical challenge to provide the rigorous existence theorem. Another interesting direction is to find some realistic applications of the model. Some progress in this direction is already obtained and I proceed to the description of it.

5. The applications

Nonlinear fields such as $\vec{n}(x)$ rarely enter the dynamical models directly. However they can appear as a part of degrees of freedom in a suitable parameterization of the original fields. For example in condensed matter theory one uses the complex valued functions $\psi_\alpha(x)$, $\alpha = 1, \ldots, N$ to describe the density amplitudes of Bose gas or the gap function of superconductor. The interaction supports the configurations, for which

$$\rho^2 = \sum_{\alpha=1}^{N} |\psi_\alpha|^2 \tag{5.7}$$

is nonvanishing. In this case it is natural to use ρ as one of the independent variables and introduce new variables

$$\chi_\alpha = \psi_\alpha/\rho$$

such that

$$\sum_{\alpha=1}^{N} |\chi_\alpha|^2 = 1. \tag{5.8}$$

In this way the compact target (I use the slang of the string theory) $\mathbb{S}^{2N-1}$ appears.

When magnetic interaction is introduced the invariance with respect to the phase transformation

$$\psi_\alpha(x) \to e^{i\lambda(x)}\psi_\alpha(x)$$

is invoked. This means that the target $\mathbb{S}^{2N-1}$ changes

$$\mathbb{S}^{2N-1} \to \mathbb{S}^{2N-1}/U(1).$$

In particular for $N = 2$ we have

$$\mathbb{S}^3/U(1) \sim \mathbb{S}^2$$

and the field $\vec{n}(x)$ naturally appears. Quite satisfactorily the tensor H_{ik} also emerges as a contribution to the magnetic field strength.

Let us illustrate it in more detail. From the beginning we shall treat the stationary system, so no electric field will be used.

The magnetic field is described in a usual way by means of the vector potential $A_k(x)$ and its interaction with ψ-fields is introduced via covariant derivatives

$$\nabla_k \psi = \partial_k \psi + i A_k \psi.$$

The energy density (of Landau-Ginsburg-Gross-Pitaevsky type) looks as follows

$$E = \sum_{\alpha=1}^{2} |\nabla_k \psi_\alpha|^2 + \frac{1}{2} F_{ik}^2 + V(|\psi_\alpha|), \tag{5.9}$$

where

$$F_{ik} = \partial_i A_k - \partial_k A_i$$

is the field strength of the magnetic field. The energy is invariant with respect to the gauge transformations

$$\psi_\alpha \to e^{i\lambda} \psi_\alpha, \quad A_k \to A_k - \partial_k \lambda.$$

We shall make the change of the field variables so that only gauge invariant ones will remain. For that observe that the first term in the RHS of (5.9) is a quadratic form in A

$$\sum_{\alpha=1}^{2} |\nabla_k \psi_\alpha|^2 = \sum_{\alpha=1}^{2} |\partial_k \psi_\alpha|^2 + A_k J_k + \rho^2 A_k^2,$$

where we use variable ρ from (5.7) and introduce current

$$J_k = -i \sum_{\alpha=1}^{2} (\bar{\psi}_\alpha \partial_k \psi_\alpha - \partial_k \bar{\psi}_\alpha \psi_\alpha).$$

It is easy to check, that under the gauge transformations the current J_k changes as follows

$$J_k \to J_k + 2\rho^2 \partial_k \lambda,$$

so that the sum

$$C_k = A_k + \frac{1}{2\rho^2} J_k$$

is gauge invariant. We shall use this variable instead of A_k. Another gauge invariant combination is given by the quadratic form

$$\vec{n} = (\bar{\chi}_1, \bar{\chi}_1)\, \vec{\tau} \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix}, \tag{5.10}$$

where $\vec{\tau} = (\tau_1, \tau_2, \tau_3)$ are Pauli matrices

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Normalization (5.8) for χ_α implies that $\vec{n}$ is a real unit vector. In fact the map

$$(\chi_1, \chi_2) \to \vec{n}$$

defined in (5.10) is a standard Hopf map. Variable $\vec{n}$ is manifestly gauge invariant and the set of variables $(\rho, \vec{n}, C_k)$ is our gauge invariant choice, substituting for the initial set (ψ_α, A_k). The energy density can be explicitly expressed via $\rho, \vec{n}$ and C_k as follows

$$E = (\partial_k \rho)^2 + \rho^2((\partial_k \vec{n})^2 + C_k^2) + \frac{1}{2}(\partial_k C_i - \partial_i C_k + H_{ik})^2 + v(\rho, n_3).$$

The most notable feature is the appearance of the tensor H_{ik}, defined in (2.3). The model, described in the main text, emerges if we put $\rho = \text{const}$ and $C = 0$. Hopefully nontrivial ρ and C, at least confined to some range, do not spoil the soliton picture. This problem is under discussion now, see [13], [18].

Let us stress, that the use of two fields ψ_α, $\alpha = 1, 2$ is most essential in this example. If $N = 1$ only variables ρ and C remain after the reduction, similar to just described. If $N > 2$ the $CP(N-1)$ field generalizing $\vec{n}$ has no topological characteristics.

Another application, considered recently [12], deals with the parameterization for the $SU(2)$ Yang-Mills field $A_\mu^a(x)$, $\mu = 0, 1, 2, 3$, $a = 1, 2, 3$. The Yang-Mills Lagrangian is invariant with respect to the nonabelian gauge transformations

$$\delta A_\mu^a = \partial_\mu \varepsilon^a + f^{abc} A_\mu^b \varepsilon^c.$$

However in some treatments one reduces this invariance by the partial gauge fixing to the abelian one

$$\delta B_\mu = i\varepsilon B_\mu\,, \quad \delta A_\mu^3 = \partial_\mu \varepsilon,$$

where $B_\mu = A_\mu^1 + iA_\mu^2$ is a complex vector field. I shall not discuss the reason for this reduction here and proceed assuming that it is done. Observe, that two vector fields A_μ^1, A_μ^2 in generic situation define a plane in Minkowski space and introduce an orthonormalized basis in this plane e_μ^α, $\alpha = 1, 2$.

$$e_\mu^\alpha e_\mu^\beta = \delta_{\alpha\beta}.$$

Let $e_\mu = e_\mu^1 + ie_\mu^2$. The basis is defined up to rotation

$$e_\mu \to e^{i\omega} e_\mu.$$

The fields B_μ can be written in terms of this basis as

$$B_\mu = \psi_1 e_\mu + \psi_2 \bar{e}_\mu$$

and thus two complex valued fields ψ_1 and ψ_2 appear. The situation becomes quite similar to the previous example and indeed in [12] the complete parameterization of

the Yang-Mills variables is introduced with appearance of $\vec{n}$-field and corresponding H-tensor. This is an indication that the Yang-Mills theory can have string-like excitations. However the situation is not that simple. The classical Yang-Mills theory is conformally invariant and has no dimensional parameters. Thus no hope for the localized regular classical solution exists. Nevertheless this complication could be lifted by quantum corrections. The famous "dimensional transmutation", which leads to the appearance of dimensional parameter in quantum effective action, could favor the nonvanishing value of the corresponding ρ-variable. All these considerations at the moment are rather speculative and need much more work to become reasonable. Personally I am quite impressed by this possibility and continue to work on it.

6. Conclusions

I think that the topic of my talk is quite instructive. It connects different domains in mathematics and mathematical physics: nonlinear PDE, elementary topology, quantum field theory, numerical methods. It illustrates the essential unity of mathematics, theoretical and applied. Finally it could lead to the realistic physical applications. For all these reasons I decided to present it to the ICM2002.

References

[1] M. D. Kruskal, N. Zabusky, Interaction of "Solitons" in a Collisionless Plasma and the Recurrence of the Initial States. *Phys. Rev. Letters*, **15** (1965), 240–243.

[2] G. S. Gardner, J. M. Greene, M. D. Kruskal, R. M. Miura, Method for solving the Korteveg-de-Vries equation. *Phys. Rev. Lett.*, **19** (1967), 1095.

[3] V. E. Zakharov, A. B. Shabat, Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, *Soviet Phys. JETP*, **34** (1972), 62–69.

[4] L. D. Faddeev, L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons*, Springer-Verlag Berlin Hiedelberg 1987.

[5] L. D. Faddeev, *How algebraic Bethe Ansatz works for integrable models*, Proc. of Les Houches summer school, session LXIV, 149–220, NATO ASI, Elsevier 1998.

[6] L. D. Faddeev, Quantization of Solitons. *Preprint IAS* print-75-QS70, 1975.

[7] T. H. R. Skyrme, A Nonlinear Field Theory, *Proc. Roy. Soc. London*, **A260** (1969), 127–138.

[8] G. 't Hooft, Magnetic Monopoles in Unified Gauge Theories, *Nucl.Phys.*, **B79** (1974), 276–284.
A. M. Polyakov, Particle Spectrum in the Quantum Field Theory. *Pisma Zh.Eksp.Teor.Fiz.*, **20** (1974), 430–433 (in Russian), *JETP Lett.* **20** (1974), 194–195.

[9] L. D. Faddeev, A. Niemi, Knots and Particles, *Nature*, **387** (1997), 58.

[10] J. Hietarinta, P. Salo, Faddeev-Hopf Knots: Dynamics of Linked Unknots. Phys. Lett., **B451** (1999), 60–67.

[11] R. Battye, P. M. Sutcliffe, Knots as Stable Soliton Solutions in a Three-Dimensional Classical Field Theory. *Phys. Rev. Lett.*, **81** (1998), 4798–4801.
[12] L. D. Faddeev, A. Niemi, Aspects of Electric-Magnetic Duality in SU(2) Yang-Mills Theory. *Phys. Lett.*, **B525** (2002), 195–200.
[13] E. Babaev, L. D. Faddeev, A. Niemi, Hidden Symmetry and Duality in a Charged Two Condensate Bose System. *Phys. Rev.*, **B65** (2002), 100512.
[14] M. I. Monastyrsky, *Topology of Gauge Fields and Condensed Matter*, Plenum, New York, USA, 1993.
[15] A. F. Vakulenko, L. V. Kapitansky, Stability of Solitons in S^2 Nonlinear σ-model, *Doklady of Soviet Acad. Sci.*, **246** (1979), 840 (in Russian), English translation in *Sov. Phys. Dokl.*, **24** (1979), 433.
[16] Yu. P. Rybakov, Structure of Minimizators of Energy in S^2 Nonlinear σ-model. *Vestnik of Lumumba Univ. (PUDN), sec. "Mathematika"* **2** (1995) 35–41.
[17] J. Hietarinta, See page in http://users.utu.fi/hietarin/knots/index.html
[18] A. P. Protogenov, Charge Density Bounds in Superconducting States of Strongly Correlated Systems, *e-Print Archive:* cond-mat/0205133.

ICM 2002 · Vol. I · 245–272

Mathematical Foundations of Modern Cryptography: Computational Complexity Perspective

Shafi Goldwasser*

Abstract

Theoretical computer science has found fertile ground in many areas of mathematics. The approach has been to consider classical problems through the prism of computational complexity, where the number of basic computational steps taken to solve a problem is the crucial qualitative parameter. This new approach has led to a sequence of advances, in setting and solving new mathematical challenges as well as in harnessing discrete mathematics to the task of solving real-world problems.

In this talk, I will survey the development of modern cryptography — the mathematics behind secret communications and protocols — in this light. I will describe the complexity theoretic foundations underlying the cryptographic tasks of encryption, pseudo-randomness number generators and functions, zero knowledge interactive proofs, and multi-party secure protocols. I will attempt to highlight the paradigms and proof techniques which unify these foundations, and which have made their way into the mainstream of complexity theory.

2000 Mathematics Subject Classification: 68Qxx, 11xx.
Keywords and Phrases: Crytography, complexity theory, One-way functions, Pseudo randomness, Computational indistinguishability, Zero knowledge interactive proofs.

1. Introduction

The mathematics of cryptography is driven by real world applications. The original and most basic application is the wish to communicate privately in the presence of an eavesdropper who is listening in. With the rise of computers as means of communication, abundant other application arise, ranging from verifying

*Department of Computer Science and Applied Mathematics, Weizmann Institute, Israel and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA. E-mail: shafi@theory.lcs.mit.edu

authenticity of data and access priveleges to enabling complex financial transactions over the internet involving several parties each with its own confidential information.

As a rule, in theoretical fields inspired by applications, there is always a subtle (and sometimes not so subtle) tension between those who do "theory" and those who "practice". At times, the practitioner shruggs of the search for a provably good method, saying that in practice his method works and will perform much better when put to the test than anything for which a theorem could be proved. The theory of Cryptography is unusual in this respect. Without theorems that provably guarantee the security of a system, it is in a sense worthless, as there is no observable outcome of using a security system other than the guarantee that no one will be able to crack it.

In computational complexity based cryptography one takes feasible (or easy) to mean those computations that terminate in polynomial time and infeasible (or hard) those computations that do not[1]. Achieving many tasks of cryptography relies on a gap between feasible algorithms used by the legitimate user versus the infeasibility faced by the adversary. On close examination then, it becomes apparent that a necessary condition for many modern cryptographic goals is that $NP \neq P$ [2], although it is not known to be a sufficient condition. A (likely) stronger necessary condition which is also sufficient for many tasks is the existence of *one-way functions*: those functions which are easy to compute but hard to invert with non-negligible probability of success taken over a polynomial time samplable distribution of inputs.

In 1976 when Diffie and Hellman came out with their paper "New Direction in Cryptography" [20] announcing that we are "on the brink of a revolution in cryptoghraphy" hopes were high that the resolution of the celebrate P vs. NP problem was close at hand and with it techniques to lower bound the number of steps required to break cryptosystems. That did not turn out to be the case. As of today, no non-linear lower bounds are known for any NP complete problem[3].

Instead, we follow a 2-step program when faced with a cryptographic task which can not be proved unconditionally (1) find the minimal assumptions necessary and sufficient for the task at hand. (2) design a cryptographic system for the task and prove its security if and only if the minimal assumptions hold. Proofs of security then are realy proofs of secure design. They take a form of a constructive reduction. For example, the existence of a one-way function has been shown a sufficient and necessary condition for "secure" digital signatures to exist[29, 52, 60]. To prove this statement one must show how to convert any "break" of the digital signature scheme into an efficient algorithm to invert the underlying one-way function. Defining formally "secure" and "break" is an essential preliminary step in accomplishing this program.

[1] We remark however that all security definitions (although not necessarily all security proofs) still make sense for a different meaning of 'easy' and 'hard'. For example, one may take easy to mean linear time whereas hard to mean quadratic time.)

[2] This is the celebrated unresolved NP vs. P problem posed by Karp, Cook and Levin in the early seventies. NP corresponds to those problems for which given a solution its correctness be verified in polynomial time whereas P corresponds to those problems for which a solution can be found in polynomial time.

[3] NP-complete problems are the hardest problems for NP. Namely, if an NP complete problem can be solved in polynomial time and thus be in P, then all problems in NP are in P.

These type of constructive reductions are a double edged sword. Say that system has been proved secure if and only if integer factorization is not in polynomial time. Then, either the system is breakable and then the reduction proof immediately yields a polynomial time integer factorization algorithm which will please the mathematicians to no end, or there exists no polynomial time integer factorization algorithms and we have found a superb cryptosystem with guaranteed security which will please the computer users to no end.

Curiously, whereas early hopes of complexity theory producing lower bounds have not materialized, cryptographic research has yielded many dividends to complexity theory. New research themes and paradigms, as well as techniques originating in cryptography, have made their way to the main stream of complexity theory. Well known techniques include random self-reducibility, hardness amplification, low degree polynomial representations of Boolean functions, and proofs by hybrid and simulation arguments. Well known examples of research themes include : interactive and probabilisticly checkable proofs and their application to show inapproximability of NP-hard algorithmic problems, the study of average versus worst case hardness of functions, and trading off hardness of computation for randomness to be used for derandomizing probabilistic complexity classes.

These examples seem, on a superficial level, quite different from each other. There are similarities however, in addition to the fact that they are investigated by a common community of researchers, who use a common collection of techniques. In all of the above, an "observer" is always present, success and failure are defined "relative to the observer", and if the observer cannot "distinguish" between two probabilistic events, they are treated as identical. This is best illustrated by examples. (1) A probabilistically checkable proofs is defined to achieve soundness if the process of checking it errs with exponentially small probability (which is indistinguishable from zero). (2) A function is considered hard to compute if all observers fail to compute it with non negligible probability taken over a efficiently samplable input distribution. It is not considered "hard" enough if it is only hard to compute with respect to some worst case input never to be encountered by the observer. (3) A source outputting bits according to some distribution is defined as pseudorandom if no observer can distinguish it from a truly random source (informally viewed as an on going process of flipping a fair coin).

1.1. Cryptography and classical mathematics

Computational infeasibility, which by algorithmic standards is the enemy of progress, is actually the cryptographer's best friend. When a computationally difficult problem comes along with some additional properties to be elaborated on in this article, it allows us to design methods which while achieving their intended functionality are "infeasible" to break. Luckily, such computationally intensive problems are abundant in mathematics. Famous examples include *integer factorization, finding short vectors in an integer lattice, and elliptic curve logarithm problem.* Viewed this way, cryptography is an external customer of number theory, algebra, and geometry. However, the complexity theory view point has not left these fields untouched, and often shed new light on old problems.

In particular, the history of cryptography and complexity theory is intertwined with the development of algorithmic number theory. This is most evident in the invention of faster tests for integer primality testing and integer factorization [48] whose quality is attested by complexity analysis rather than the earlier benchmarking of their performance. A beautiful account on the symbiotic relationship between number theory and complexity theory is given by Adleman [2] who prefaces his article by saying that "Though algorithmic number theory is one of man's oldest intellectual pursuit, its current vitality is unrivaled in history. This is due in part to the injection of new ideas from computational complexity."

1.2. Cryptography and information theory

In a companion paper to his famous paper on information theory, Shannon [66] introduced a rigorous theory of perfect secrecy based on information theory. The theory addresses adversary algorithms which have unlimited computational resources. Thus, all definitions of security, which we will refer to henceforth as *information theoretic security*, and proofs of possibility and impossibility are with respect to such adversary. Shannon proves that "perfectly secure encryption" can only exist if the size of secret information that legitimate parties exchange between them in person prior to remote transmission, is as large as the total entropy of secret messages they exchange remotely. Maurer [51] generalized these bounds to two-way communications. This limits the practice of encryption based on information theory a great deal. Even worse, the modern cryptographic tasks of public-key encryption, digital signatures, pseudo random number generation, and most two party protocols can be proved down right impossible information theoretically. To achieve those, we turn to adversaries who are limited computationally and aim at computational security with the cost of making computational assumptions or assumptions about the physical world.

Having said that, some cryptographic tasks can achieve full information theoretic security. A stellar example is of multi party computation. Efficient and information theoretic secure multi-party protocols are possible unconditionally tolerating less than half faults, if there are perfect private channels between each pair of honest users [8, 19, 61, 33]. Statistical *zero-knowledge proofs* are another example [32, 71].

Perfect private channels between pairs of honest users can be implemented in several settings: (1) The *noisy channel* setting [45] (which is a generalization of the *wire tal channel* [75]) where the communication between users in the protocol as well as what the adversary taps is subject to noise). (2) A setting where the adversary's memory (i.e. ability to store data) is limited [18]. (3) The *Quantum Channels* setting where by quantum mechanics, it is impossible for the adversary to obtain full information on messages exchanged between honest users. Introducing new and reasonable such settings which enable information theoretic security is an important activity.

Moreover, often paradigms and construction introduced within the computational security framework can be and have been lifted out to achieve information theoretic security. The development of randomness extractors from pseudo random

number generators can be done in this fashion [72].

We note that whereas the computational complexity notions of secrecy, knowledge, and pseudo-randomness are different than their information theoretic analogues, techniques of error recovery developed in information theory are extremely useful. Examples include the Haddamard error correcting codes which is used to exhibit hard core predicates in one-way functions [28], and various polynomial based error correcting codes which enable high fault tolerance in multi-party computation [8].

To sum up, the theory of cryptography has in the last 30 years turned into a rich field with its own rules, structure, and mathematical beauty which has helped to shape complexity theory. In the talk, I will attempt to lead you through a short summary of what I believe to have been a fascinating journey of modern cryptography. I apologize in advance for describing my own journey, at the expense of other points of view. I attach a list of references including several survey articles that contain full details and proofs [40].

In the rest of the article, I will briefly reflect on a few points which will make my lecture easier to follow.

2. Conventions and complexity theory terminology

We say that an algorithm is *polynomial time* if for all inputs x, the algorithm runs in time bounded by some polynomial in $|x|$ where the latter denotes the length of x when represented as a binary string. A *probabilistic algorithm* is one that can make random choices, where without loss of generality each choice is among two and is taken with probability $1/2$. We view these choices as the algorithm *coin tosses.* A probabilistic algorithm A on input x may have more than one possible output depending on the outcome of its coin tosses, and we will let $A(x)$ denote the probability distribution over all possible outputs. We say that a probabilistic algorithm is *probabilistic polynomial time* (PPT) if for any input x, the expectation of the running time taken over the all possible coin tosses is bounded by some polynomial in $|x|$, regardless of the outcome of the coin tosses.

In complexity theory, we often speak of language classes. A language is a subset of all binary strings. The class P is the set of languages such that there exists a polynomial time algorithm, which on every input x can decide if x is in the language or not. The class BPP are those languages whose membership can be decided by a probabilistic polynomial time algorithm which for every input, is incorrect with at most negligible probability taken over the coin tosses of the algorithm. The class NP is the class of languages accepted by polynomial time non-deterministic algorithm which may make non-deterministic choices at every point of computation. Another characterization of NP is as the class of languages that have short proofs of memberships. Formally, $NP = \{L|$ there exists polynomial time computable function f and $k > 0$, such that $x \in L$ iff there exists y such that $f(x, y) = 1$ and $|y| < |x|^k\}$.

In this article, we consider an 'easy' computation to be one which is carried out by a PPT algorithm. A function ν: $\mathbf{N} \rightarrow \mathbf{R}$ is negligible if it vanishes faster than the inverse of any polynomial. All probabilities are defined with respect to finite probability spaces.

3. Indistinguishability

Indistinguishability of probability distributions is a central concept in modern cryptography. It was first introduced in the context of defining security of encryption systems by Goldwasser and Micali [31]. Subsequently, it turned out to play a fundamental role in defining pseudo-randomness by Yao [76], and zero-knowledge proofs by Goldwasser, Micali, and Rackoff [32].

Definition 1 *Let $X = \{X_k\}_k$, $Y = \{Y_k\}$ be two ensembles of probability distributions on $\{0,1\}^k$. We say that X is* **computationally indistinguishable** *from Y if $\forall$ probabilistic polynomial time algorithms A, $\forall\ c > 0$, $\exists k_0$, $s.t\ \forall k > k_0$,*

$$|\Pr_{t \in X_k}(A(t) = 1) - \Pr_{t \in Y_k}(A(t) = 1)| < \frac{1}{k^c}.$$

The algorithm A used in the above definition is called a polynomial time statistical test.

Namely, for sufficiently long strings, no probabilistic polynomial time algorithms can tell whether the string was sampled according to X or according to Y. Note that such a definition cannot make sense for a single string, as it can be drawn from either distribution. Although we chose to focus on polynomial time indistinguishability, one could instead talk of distribution which are indistinguishable with respect to any other computational resource, in which case all the algorithms A in the definition should be bounded by the relevant computational resource. This, has been quite useful when applied to space bounded computations [53].

Of particular interest are those probability distributions which are indistinguishable from the uniform distribution, focused on in [76], and are called *pseudo-random distributions.*

Let $U = \{U_k\}$ denote the uniform probability distribution on $\{0,1\}^k$. That is, for every $\alpha \in \{0,1\}^k$, $\Pr_{x \in U_k}[x = \alpha] = \frac{1}{2^k}$.

Definition 2 *We say that $X = \{X_k\}_k$ is* **pseudo random** *if it is computationally indistinguishable from U. That is, $\forall$ probabilistic polynomial time algorithms A, $\forall\ c > 0\ \exists k_0$, such that $\forall k > k_0$,*

$$|\Pr_{t \in X_k}[A(t) = 1] - \Pr_{t \in U_k}[A(t) = 1]| < \frac{1}{k^c}.$$

If $\exists A$ and c such that the condition in definition 2 is violated, we say that X_k fails the statistical test A.

A simple but not very interesting example of two probability distributions which are computationally indistinguishable are two distributions which are statistically very close. For example, $X = \{X_k\}$ defined exactly as the uniform distribution over $\{0,1\}^k$ with two exceptions, 0^k appears with probability $\frac{1}{2^{k+1}}$ and 1^k appears with probability $\frac{3}{2^{k+1}}$. Then the uniform distribution and X can not be distinguished by any algorithm (even one with no computational restrictions) as long as it is only given a polynomial size sample from one of the two distributions.

It is fair to ask as this point whether computationally indistinguishability is anything more than statistical closeness where the latter is formally defined as follows.

Definition 3 *Two probability distributions X, Y are statistically close if $\forall c > 0$, $\exists k_0$ such that $\forall k > k_0$,*

$$\sum_t |\Pr(t \in X_k) - \sum_t (t \in U_k) < \frac{1}{k^c}.$$

X and Y are far if they are not close.

Do there exist distributions which are statistically far apart and yet are computationally indistinguishable? Goldreich and Krawczyk [27] who pose the question note this to be the case by a counting argument. However their argument is non constructive. The works on secure encryption and pseudo random number generators [31, 10, 76] imply the existence of *efficiently constructible* pairs of distributions that are computationally indistinguishable but statistically far, under the existence of one-way functions. The use of assumptions is no accident.

Theorem 4 [25] *The existence of one-way functions is equivalent to the existence of pairs of polynomial-time constructible distributions which are computationally indistinguishable and statistically far.*

4. Building blocks

A central building block required for many tasks in cryptography is the existence of a one-way function. Let us discuss this basic primitive as well as a few others in some detail.

4.1. One-way functions

Informally, a one-way function is a function which is "easy" to compute but "hard" to invert. Any probabilistic polynomial time (PPT) algorithm attempting to invert the function on an element in its range, should succeed with no more than "negligible" probability, where the probability is taken over the elements in the domain of the function and the coin tosses of the PPT attempting the inversion. We often refer to an algorithm attempting to invert the function as an adversary algorithm.

Definition 5 *A function* f: $\{0,1\}^* \to \{0,1\}^*$ *is* one-way *if:*

1. *Easy to Evaluate: there exists a PPT algorithm that on input* x *output* $f(x)$*;*
2. *Hard to Invert: for all PPT algorithm* A*, for all* $c > 0$*, there exists* k_0 *such that for all* $k > k_0$*,*

$$\Pr\left[A(1^k, f(x)) = z \ : \ f(x) = f(z)\right] \leq \frac{1}{k^c}$$

where the probability is taken over $x \in \{0,1\}^k$ *and the coin tosses of* A*.*

Note Unless otherwise mentioned, the probabilities during this section are calculated uniformly over all coin tosses made by the algorithm in question.

A few remarks are in order. (1)The guarantee is probabilistic. The adversary has low probability of inverting the function where the probability distribution is taken over the inputs of length k to the one-way function and the possible coin tosses of the adversary.

(2) The adversary is not asked to find x; that would be pretty near impossible. It is asked to find some inverse of $f(x)$. Naturally, if the function is 1-1 then the only inverse is x. We note that it is much easier to find candidate one-way functions without imposing further restrictions on its structure, but being 1-1 or at least *regular* (that is, the number of preimage of any image is about of the range), it results in easier and more efficient cryptographic constructions.

(3) One may consider a non-uniform version of the "Hard to invert" requirement, requiring the function to be hard to invert by all non-uniform polynomial size family of algorithms, rather than by all probabilistic polynomial time algorithms. The former extends probabilistic polynomial time algorithms to allow for each different input size, a different polynomial size algorithm.

(4) The definition is typical to definitions from computational complexity theory, which work with asymptotic complexity—what happens as the size of the problem becomes large. One-wayness is only asked to hold for large enough input lengths, as k goes to infinity. Per this definition, it may be entirely feasible to invert f on, say, 512 bit inputs. Thus such definitions are useful for studying things on a basic level, but need to be adapted to be directly relevant to practice.

(5) The above definition can be considerably weakened by replacing the second requirement of the function to require it to be hard to invert on **some** non-negligible fraction of its inputs (rather than all but non-negligible fraction of its inputs). This relaxation to a *weak one-way function* is motivated by the following example. Consider the function $f : \mathbf{Z} \times \mathbf{Z} \mapsto \mathbf{Z}$ where $f(x,y) = x \cdot y$. This function can be easily inverted on at least half of its outputs (namely, on the even integers) and thus is not a one-way function as defined above. Still, f resists all efficient algorithms when x and y are primes of roughly the same length which is the case for a non-negligible fraction ($\approx \frac{1}{k^2}$) of the k-bit composite integers. Thus according to our current state of knowledge of integer factorization, f does satisfy the weaker requirement. Convertion between any weak one-way function to a one-way function have been shown using "hardness amplification" techniques which expand the size of the input by a polynomial factor [76]. Using expanders, constant factor expansions (of the input size) construction of a one-way function from a weak one-way function

is possible [26].

(6) To apply this definition to practice we must typically envisage not a single one-way function but a family of them, parameterized by a *security parameter* k. That is, for each value of the security parameter k, there is a family of functions, each defined over some finite domain and finite ranges. The existence of a single one-way function is equivalent to the existence of a collection of one-way functions.

Definition 6 *A collection of one-way functions is a set* $F = \{f_i : D_i \to R_i\}_{i \in I}$ *where* I *is an index set, and* D_i *(*R_i*) are finite domain(range) for* $i \in I$*, satisfying the following conditions.*

1. *Selection in Collection:* $\exists$ *PPT algorithm* S_1 *that on input* 1^k *outputs an* $i \in I$ *where* $|i| = k$.
2. *Selection in Domain:* $\exists$ *PPT algorithm* S_2 *that on input* $i \in I$ *outputs* $x \in D_i$
3. *Easy to Evaluate:* $\exists$ *PPT algorithm Eval such that for* $i \in I$ *and* $x \in D_i$, $Eval(i, x) = f_i(x)$.
4. *Hardness to Invert:* $\forall$ *PPT adversary algorithm* A, $c > 0$, $\exists$ k_0 *such that* $\forall$ $k > k_0$,

$$\Pr\left[A(1^k, i, f_i(x)) = z \ : \ f(x) = f(z)\right] \leq \frac{1}{k^c}$$

(the probability is taken over $i \in S_1(1^k), x \in S_2(i)$ *and the coin tosses of* A*).*

The hardness to invert condition can be made weaker by requiring only that $\exists c > 0$, such that $\forall$ PPT algorithm A, $\exists$ k_0 such that $\forall$ $k > k_0$. $Prob[A(1^k, i, f_i(x)) \neq z,\ f(x) = f(z)] > \frac{1}{k^c}$ (the probability taken over $i \in S_1(1^k), x \in S_2(i)$ and the coin tosses of A). We call collections which satisfy such weaker conditions, collection of weak one-way functions. Transformations exist via sampling algorithms between both types of collections.

Another useful and equivalent notion is of a *one-way predicate*, first introduced in [31]. This is a Boolean function of great use in encryption and protocol design. A one-way predicate is equivalent to the existence of 0/1 problems, for which it is possible to uniformly select an instance for which the answer is 0 (or respectively 1), and yet for a (pre-selected) instance it is hard to compute with success probability greater than $\frac{1}{2}$ whether the answer is 0 or 1.

Definition 7 *A one-way predicate is a Boolean function* $B : \{0,1\}^* \to \{0,1\}$ *for which*

1. *Sampling is possible:* $\exists$ *PPT algorithm* S *that on input* $v \in \{0,1\}$ *and* 1^k*, outputs a random* x *such that* $B(x) = v$ *and* $x \in \{0,1\}^k$.
2. *Guessing is hard:* $\forall c > 0$, $\forall$ *PPT algorithms* A, $\forall k$ *sufficiently large,* $Prob[A(x) = B(x)] \leq \frac{1}{2} + \frac{1}{k^c}$ *(probability is taken over* $v \in \{0,1\}, x \in S(1^k, v)$*, and the coin tosses of* A*).*

Proving the equivalence between one-way predicates and one-way functions is easy in the forward direction, by viewing the sampling algorithm S as a function over its coin tosses. To prove the reverse implication is quite involved. Toward this goal, the notion of a hard core predicate of a one-way function was introduced in [10, 76]. Jumping ahead, hard core predicate of one-way functions yield immediately one-way predicates.

4.1.1. Hard-core predicates

The fact that f is a one-way function obviously does not necessarily imply that $f(x)$ hides everything about x. It is easy to come up with constructions of universal one-way functions in which one of the bits of x *leaks* from $f(x)$. Even if each bit of x is well hidden by $f(x)$ then some function of all of the bits of x can be easy to compute. For example, the least significant bit of x is easy to compute from $f_{p,g}(x) = g^x \bmod p$ where p is a prime and g a generator for the cyclic group Z_p^*, even though we know of no polynomial time algorithms to compute x from $f_{p,g}(x)$. Similarily, it is easy of compute the Jacobi symbol of x mod n from the RSA function $RSA_{n,e}(x) = x^e \bmod n$ where $(e, \phi(n)) = 1$, even though the fastest algorithm to invert $RSA_{n,e}$ needs to factor integer n first, which is not known to be a polynomial time computation.

Yet, clearly there are some bits of information about x which cannot be computed from $f(x)$, given that x in its entirety is hard to compute. The question is, which bits of x are hard to compute, and how hard to compute are they. The answer is encouraging. For several functions f for which no polynomial time inverting algorithm is known, we can identify particular bits of the pre-image of f which can be proven (via a polynomial time reduction) to be as hard as to compute with probability significantly better than $\frac{1}{2}$, as it is to invert f itself in polynomial time. Examples of these can be found in [10, 31, 36, 1].

More generally, a *hard-core predicate* for f, is a Boolean predicate about x which is efficiently computable given x, but is hard to compute from $f(x)$ with probability significantly better than $\frac{1}{2}$.

Definition 8 *A hard-core predicate of a function $f : \{0,1\}^* \to \{0,1\}^*$ is a Boolean predicate $B : \{0,1\}^* \to \{0,1\}$, such that*

1. *$\exists PPT$ algorithm $Eval$, such that $\forall x\, Eval(x) = B(x)$*
2. *$\forall$ PPT algorithm A, $\forall c > 0$, $\exists k_0$, s.t. $\forall k > k_0$ $\Pr[A(f(x)) = B(x)] < \frac{1}{2} + \frac{1}{k^c}$. The probability is taken over the random coin tosses of A, and random choices of x of length k.*

Yao proposed a construction of a hard-core predicate for any one-way function [76]. A considerably simpler construction and proof general result is due to Goldreich and Levin [28].

Theorem 9 [28] *Let f be a length preserving one-way function. Define $f'(x \circ r) = f(x) \circ r$, where $|x| = |r| = k$, and $\circ$ is the concatenation function. Then*

$$B(x \circ r) = \Sigma_{i=1}^{k} x_i r_i (mod\ 2)$$

is a hard-core predicate for f' (Notice that if f is one-way then so is f').

Interestingly, the proof of the theorem can be regarded as the first example of a polynomial time list decoding [63] algorithm. Essentially $B(x,r)$ may be viewed as the rth bit of a Haddamrd encoding of x. The proof of the theorem yields a polynomial time error decoding algorithm which returns a polynomial size list of candidates for x, as long as the encoding is subject to an error rate of less than $\frac{1}{2} - \epsilon$ where $\epsilon > \frac{1}{k^c}$ for some constant $c > 0$, $k = |x|$. The length of the list is $O(\frac{1}{\epsilon^2})$.

4.2. Trapdoor functions

A *trapdoor function* f is a one-way function with an extra property. There also exists a secret inverse function (the *trapdoor*) that allows its possessor to efficiently invert f at any point in the domain of his choosing. It should be easy to compute f on any point, but infeasible to invert f with high probability without knowledge of the inverse function. Moreover, it should be easy to generate matched pairs of f's and corresponding trapdoor.

Definition 10 *A trapdoor function is a one-way function* $f : \{0,1\}^* \to \{0,1\}^*$ *such that there exists a polynomial* p *and a probabilistic polynomial time algorithm* I *such that for every* k *there exists a* $t_k \in \{0,1\}^*$ *such that* $|t_k| \leq p(k)$ *and for all* $x \in \{0,1\}^k$, $I(f(x), t_k) = y$ *such that* $f(y) = f(x)$.

Trapdoor functions are much harder to locate than one-way function, as they seem to require much more hidden structure. An important problem is to establish whether one implies the other. Recent results of [41] indicate this may not the case.

A *trapdoor predicate* is a one-way predicate with an extra trapdoor property: for every k, there must exist trapdoor information t_k whose size is bounded by a polynomial in k and whose knowledge enables the polynomial-time computation of $B(x)$, for all $x \in \{0,1\}^k$. Restating as a collection of trapdoor predicates we get.

Definition 11 *Let* I *be an index set and for* $i \in I$, D_i *a finite domain. A collection of trapdoor is a set* $B = \{B_i : D_i \to \{0,1\}\}_{i \in I}$ *such that:*

1. $\exists$ *PPT algorithm* S_1 *which on input* 1^k *outputs* (i, t_i) *where* $i \in I \cap \{0,1\}^k$, *and* $|t_i| < poly(k)$ *(* t_i *is the trapdoor).*
2. $\exists$ *PPT algorithm* S_2 *which on input* $i \in I, v \in \{0,1\}$ *outputs* $x \in D_i$ *such that* $B_i(x) = v$.
3. $\exists$ *PPT algorithm* S_3 *which on input* $i \in I, x \in D_i, t_i$ *outputs* $B_i(x)$.
4. $\forall$ *PPT adversary algorithms* A, $c > 0$, $\exists k_0, \forall k > k_0$, $Prob[A(i,x) = B_i(x)] \leq \frac{1}{2} + \frac{1}{k^c}$ *(the probability taken over* $i \in S_1(1^k), v \in \{0,1\}, x \in S_2(i,v)$, *and the coins of* A*).*

The existence of a trapdoor predicate is equivalent to the existence of secure public-key encryption as we shall see in the next section. Trapdoor functions imply trapdoor predicates, but it is an open problem to show that they are equivalent.

Claim 12 *If trapdoor functions exist then collection of trapdoor predicates exist.*

4.3. Candidate examples of building blocks

It has been shown by a fairly straightforward diagonalization argument [39] how to construct a *universal* one-way function (i.e. a function which is one-way if any one-way function exists). Still this is very inefficient, and concrete proposals for one-way function are needed for any practical usage of cryptographic constructions which utilized one-way functions. Moreover, looking into the algebraic, combinatorial, and geometric structure of concrete proposals has lead to many insights about

what could be true about general one-way functions. The revelation process seems almost always to start from proving properties about concrete examples to generalizing to proving properties on general one-way functions.

Interesting proposals for one-way functions, trapdoor functions, and trapdoor predicates have been based on hard computational problems from number theory, coding theory, algebraic geometry, and geometry of numbers. What makes a computational problem a "suitable" candidate? First, it should be put under extensive scrutiny by the relevant mathematical community. Second, the problem should be hard on the *average* and not only in the *worst* case. A big project in cryptography is the construction of cryptographic functions which are provably hard to break *on the average* under some *worst-case* computational complexity assumption. A central technique is to show that a problem is as hard for an *average* instance as it is for a *worst case* instance by *random self reducibility* [6]. A problem P is random self reducible if there exists a probabilistic polynomial time algorithm that maps any instance I of P to a collection of random instances of P such that given solutions to the random instances, one can efficiently obtain a solution to the original instance. Variations would allow mapping any instance of P to random instances of P'.[4]

Perhaps the most interesting problem in cryptography today is to show (or rule out) that the existence of a one-way function is equivalent to the $NP \neq BPP$.

For lack of space, we discuss in brief a few proposals.

4.3.1. Discrete logarithm problem proposal

Let p be a prime integer and g a generator for the multiplicative cyclic group $\mathcal{Z}_p^* = \{1 \leq y < p | (y,p) = 1\}$. The *discrete log problem* (DLP) is given p,g, and $y \in Z_p^*$, compute the unique x such that $1 \leq x \leq p-1$ and $y = g^x \bmod p$. The discrete log problem has been first suggested to be useful for key exchange over the public channel by Diffie and Hellman [20].

The function $DL(p,g,x) = (p,g,g^x \bmod p)$, and the corresponding collection of functions $DL = \{DL_{p,g} : Z_{p-1} \rightarrow Z_p^*, DL_{p,g}(x) = g^x \bmod p\}_{<p,g> \in I}$ where $I = \{< p,g >, p \text{ prime }, g \text{ generator}\}$ have served as proposals for a one-way function and a collection of one-way functions (respectively). On one hand, there exist efficient algorithms to select pairs of (p,g) of a given length with uniform probability [7], and to perform modulo exponentiation. On the other hand, the fastest algorithms to solve the discrete log problem is the generalized number field sieve version of the index-calculus method which runs in expected time $e^{((c+o(1))(\log p)^{\frac{1}{3}}(\log\log p)^{\frac{2}{3}})}$ (see survey [54]). Moreover, for a fixed prime p, $DL(p,g,g^x \bmod p)$ can be shown as hard to invert on the average over the $1 \leq x \leq p-1$ and g generators, as it is for every g and x.

[4] This technique was first observed and applied to the number theoretic problems of factoring, discrete log, testing quadratic residuosity, and the RSA function. In each of these problems, one could use the algebraic structure to show how to map a particular input uniformly and randomly to other inputs in such a way that the answer for the original input can be recovered from the answers for the targets of the random mapping. Showing that polynomials are randomly self reducible over finite fields was applied to the low-degree polynomial representations of Boolean functions, and has been a central and useful technique in probabilistically checkable proofs.

An important open problem is to prove that, *without fixing first the prime* p, solving the discrete log problem for an average instance (p, g, y) is hard on the average as in the worst case.

In the mid-eighties an extension of the discrete logarithm problem over prime integers, to *computing discrete logarithms over elliptic curves* was suggested by Koblitz and V. Miller (see survey [46]). The attraction is that the fastest algorithms known for computing logarithms over elliptic curves are of complexity $O(\sqrt{p})$ for finite field F_p. The main concern is that they have not been around long enough to go under extensive scrutiny, and that the intersection between the mathematical community who can offer such scrutiny and the cryptographic community is not large.

4.3.2. Shortest vector in integer lattices proposal

In a celebrated paper [4] Ajtai described a problem that is hard *on the average* if some well-known integer lattice problems are *hard to approximate in the worst case*, and demonstrated how this problem can be used to construct one-way functions. Previous worst case to average case reductions were applied to two parameter problems and the reduction was shown upon fixing one parameter (e.g. in the discrete logarithm problem random self reducibility was shown fixing the prime parameter), whereas the [4] reduction is the first which averages over all parameters.

Let V be a set of n linearly independent vectors $V = \{v_1, \cdots, v_n, v_i \in \mathcal{R}\}$. The integer lattice spanned by V is the set of all possible linear combinations of the v_i's with integer coefficients, namely $L(V) \stackrel{\text{def}}{=} \{\sum_i a_i v_i \ : \ a_i \in \mathsf{Z} \text{ for all } i\}$. We call V the basis of the lattice $L(V)$. We say that a set of vectors $L \subset \mathcal{R}^n$ is a lattice if there is a basis V such that $L = L(V)$.

Finding "short vectors" (i.e., vectors with small Euclidean norm) in lattices is a hard computational problem. There are no known efficient algorithms to find or even approximate - given an arbitrary basis of a lattice - either the shortest non-zero vector in the lattice, or another basis for the same lattice whose longest vector is as short as possible. Given an arbitrary basis B of a lattice L in $\mathbf{R}^n$, the best algorithm to approximate (up to a polynomial factor in n) the length of the shortest vector in L is the L^3 algorithm [49] which approximates these problems to within a ratio of $2^{n/2}$ in the worst case, and its improvement [64] to ratio $(1+\epsilon)^n$ for any fixed $\epsilon > 0$.

Ajtai reduced the worst-case complexity of problem (W) which is closely related the length of the shortest vector and basis in a lattice, to the average-case complexity of problem (A) (version presented here is due to Goldreich, Goldwasser, and Halevi [34]).

W : Given an arbitrary basis B of a lattice L, find a set of n linearly independent lattice vectors, whose length is at most polynomially (in n) larger than the length of the smallest set of n linearly independent lattice vectors. (The length of a set of vectors is the length of its longest vector.)

A : Let parameters $n, m, q \in \mathcal{N}$ be such that $n \log q < m \leq \frac{q}{2n^4}$ and $q = O(n^c)$ for some constant $c > 0$. Given a matrix $M \in \mathsf{Z}_q^{n \times m}$, find a vector $x \in \{-1, 0, 1\}^m, x \neq 0$ so that $Mx \equiv 0 \pmod{q}$.

Theorem 13 [4, 34] *Suppose that it is possible to solve a uniformly selected instance of Problem (A) in expected $T(n, m, q)$-time, where the expectation is taken over the choice of the instance as well as the coin-tosses of the solving algorithm. Then it is possible to solve Problem (W) in expected* $\text{poly}(|I|) \cdot T(n, \text{poly}(n), \text{poly}(n))$ *time* on every *n-dimensional instance I, where the expectation is taken over the coin-tosses of the solving algorithm.*

The construction of a candidate one-way function follows in a straight forward fashion. Let M be a random $k \times m$ matrix M with entries from Z_q, where m and q are chosen so that $k \log q < m < \frac{q}{2k^4}$ and $q = O(k^c)$ for some constant $c > 0$ (k here is the security parameter).

The one-way function candidate is then $f(M, s) = (M, Ms \bmod q = \sum_i s_i M_i \bmod q)$ where $s = s_1 s_2 \cdots s_m \in \{0,1\}^m$ and M_i is the i'th column of M. We note that this function is regular.

4.3.3. Factoring integers proposal

Consider the function $Squaring(n, x) = (n, x^2 \bmod n)$ where $n = pq$ for $p, q \in Z$ prime numbers and $x \in Z_n^*$, and the corresponding collection of functions $Squaring = \{Squaring_n(x) = x^2 \bmod n : Z_n^* \to Z_n^*, n = pq, p, q \text{ primes}, |p| = |q| = k\}_k$. This function is easy to compute without knowing the factorization of n, and is easy to invert given the factorization of n (the trapdoor) using fast square root extraction algorithms modulo prime moduli [5] and the Chinese remainder theorem. Moreover, as the primes are abundant by the prime number theorem ($\approx \frac{1}{k}$ for k-bit primes) and there exist probabilistic expected polynomial time algorithms for primality testing [30, 3], it is easy to uniformly select n, p, q of the right form.

In terms of hardness to invert, Rabin [62] has shown it as hard to invert as it is to factor n as follows. Suppose there exists a factoring algorithm A. Choose $r \in Z_n^*$ at random. Let $y = A(r^2 \bmod n)$. If $y \neq r$ or $n - r$, then let $p = gcd(r - y, n)$, else choose another r and repeat. Within expected 2 trials you should obtain p. The asymptotically proven fastest integer factorization algorithm to date is the number field sieve which runs in expected time $e^{((c+o(1))(\log n)^{\frac{1}{3}}(\log\log n)^{\frac{2}{3}})}$ [59]. The hardest input to any factoring algorithms are integers $n = pq$ which are product of two primes of similar length. Finally, for a fixed n, $Squaring(n, \cdot)$ can be shown as hard to invert on the average over $x \in Z_n^*$ as it is for any x. We remark, that integer factorization has been first proposed as a basis for a trapdoor function in the celebrated work of Rivest, Shamir and Adelman [56].

By choosing p and q to be both congruent to 3 mod 4 and restricting the domain of $Squaring_n$ to the quadratic residues mod n, this collection of functions becomes a collection of permutations proposed by Williams [74], which are especially easy to work with in many cryptographic applications.

An open problem is to prove that the difficulty of factoring integers is as hard on the average as in the worst case. In our terminology an affirmative answer would mean that $x^2 \bmod n$ is as hard to invert on the *average* over n and x, as it is for *any* n and x.

4.3.4. Quadratic residues vs. quadratic non residues proposal

Let $n \in Z$. Then we call $y \in Z_n^*$ is a *quadratic residue mod* n iff $\exists x \in Z_n^*$ such that $y \equiv x^2 \bmod n$. Let us restrict our attention to $n = pq$ where $p = q = 3 \bmod 4$.

Selecting a random quadratic residue mod n is easy by choosing $r \in Z_n^*$ and computing $r^2 \bmod n$. Similarily, for such n, selecting a random quadratic non-residue is easy by choosing $r \in Z_n^*$ and computing $n - r^2 \bmod n$ (this is a quadratic non-residue by the property of the n's chosen).

On the other hand, deciding whether x is a quadratic residue modulo n for n composite (which is the case if and only if it is a quadratic residue modulo each of its prime factors), seems a hard computational problem. No algorithm is known other than first factoring n and then deciding whether x is a quadratic residue modulo all its prime factors. This is easy for a prime modulos by computing the Legendre symbol $(\frac{x}{p}) = x^{\frac{p-1}{2}} \bmod p$ $(= 1$ iff x is a quadratic residue mod $p)$. The Legendre symbol is generalizable to the Jacobi symbol for composite moduli $(\frac{x}{n}) = \Pi_{p^\alpha | n}(\frac{x}{p})^\alpha$ where $n = \Pi p^\alpha$. The Jacobi symbol only provides partial answer to whether x mod n is a quadratic residue or not. For $x \in J_n^{+1} = \{x \in Z_n^*, (\frac{x}{n}) = 1\}$, it gives no information.

A proposal by Goldwasser and Micali [31] for a collection of trapdoor predicates follows.

$QR = \{QR_n : J_n^{+1} \to \{0,1\}\}_{n \in I}$ where $I = \{n = pq || p, q,$ primes$, |p| = |q|\}$,

$$QR_n(x) = \left\{ \begin{array}{l} 0 \text{ if } x \text{ is a quadratic residue mod } n \\ 1 \text{ if } x \text{ is a quadratic non-residue mod } n \end{array} \right\}.$$

It can be proved that for every n distinguishing between random quadratic residues and random quadratic non residues with Jacobi symbol $+1$, is as hard as solving the problem entirely in the worst case.

Theorem 14 [31] *Let $S \subset I$. If there exists a PPT algorithm which for every $n \in S$, can distinguish between quadratic residues and quadratic non-residues with non-negligible probability over $\frac{1}{2}$ (probability taken over the $x \in Z_n^*$ and the coin tosses of the distinguishing algorithm), then there exist a PPT algorithm which for every $n \in S$ and every $x \in Z_n^*$ decides whether x is a quadratic residue mod n with probability close to 1.*

5. Encryption case study

As discussed in the introduction we would like to propose cryptographic schemes for which we can prove theorems guaranteeing the security of our proposals. This task includes a definition phase, construction phase and a reduction proof which is best illustrated with an example. We choose the example of encryption.

We will address here the simplest setting of a passive adversary who can tap the public communication channels between communicating parties. We will measure the running time of the encryption, decryption, and adversary algorithms as a function of a *security parameter* k which is a parameter fixed at the time the cryptosystem is setup. We model the adversary as any probabilistic algorithm which

runs in time bounded by some polynomial in k. Similarily, the encryption and decryption algorithms designed are probabilistic and run in polynomial time in k.

5.1. Encryption: definition phase

Definition 15 *A* public-key encryption scheme *is a triple, (G, E, D), of probabilistic polynomial-time algorithms satisfying the following conditions*

1. *key generation algorithm : On input 1^k (the security parameter) algorithm G, produces a pair (e, d) where e is called the public key, and d the corresponding private key. (Notation: $(e, d) \in G(1^k)$.) We will also refer to the pair (e, d) a pair of* encryption/decryption *keys.*
2. *An encryption algorithm: Algorithm E takes as inputs encryption key e from the range of $G(1^k)$ and string $m \in \{0,1\}^k$ called the message, and produces as output string $c \in \{0,1\}^*$ called the ciphertext. (We use the notation $c \in E(e, m)$ or the shorthand $c \in E_e(m)$.) Note that as E is probabilistic, it may produce many ciphertexts per message.*
3. *A decryption algorithm: Algorithm D takes as input decryption key d from the range of $G(1^k)$, and a ciphertext c from the range of $E(e, m)$, and produces as output a string $m' \in \{0,1\}^*$, such that for every pair (e, d) in the range of $G(1^k)$, for every m, for every $c \in E(e, m)$, the prob$(D(d, c) \neq m')$ is negligible.*
4. *Furthermore, this system is "secure" (see discussion below).*

A *private-key encryption scheme* is identically defined except that $e = d$. The security definition for private-key encryption and public-key encryption are different in one aspect only, in the latter e is a public input available to the whereas in the former e is a secret not available to the adversary.

5.1.1. Defining security

Brain storming about what it means to be secure brings immediately to mind several desirable properties. Let us start with the the minimal requirement and build up.

First and foremost the private key should not be recoverable from seeing the public key. Secondly, with high probability for any message space, messages should not be entirely recovered from seeing their encrypted form and the public file. Thirdly, we may want that in fact no useful information can be computed about messages from their encrypted form. Fourthly, we do not want the adversary to be able to compute any useful facts about traffic of messages, such as recognize that two messages of identical content were sent, nor would we want her probability of successfully deciphering a message to increase if the time of delivery or relationship to previous encrypted messages were made known to her.

In short, it would be desirable for the encryption scheme to be the mathematical analogy of opaque envelopes containing a piece of paper on which the message is written. The envelopes should be such that all legal senders can fill it, but only the legal recipient can open it.

Two definitions of security attempting to capture the "opaque envelope" analogy have been proposed in the work of [31] and are in use today: computational indistinguishability and semantic security. The first definition is easy to work with whereas the second seems to be the natural extension of Shannon's perfect secrecy definition to the computational world. They are equivalent to each other as shown by [31, 67].

The first definition essentially requires that the the adversary cannot find a pair of messages m_0, m_1 for which the probability distributions over the corresponding ciphertexts is computationally distinguishable.

Definition 16 *We say that a Public Key Cryptosystem (G, E, D) is computationally indistinguishable if $\forall$ PPT algorithms F, A, and for $\forall$ constant $c > 0$, $\exists k_0$, $\forall k > k_0$, $\forall m_0, m_1 \in F(1^k)$, $|m_0| = |m_1|$,*

$$|\Pr[A(e,c) = 1 \text{ where } (e,d) \in G(1^k);\, c \in E(e,m_0)] - \Pr[A(e,c) = 1 (e,d) \in G(1^k);\, c \in E(e,m_1)]| < \frac{1}{k^c}.$$

Remarks about the definition

1. In the case of private-key cryptosystem, the definition changes slightly. The encryption key e is not given to algorithm A.
2. Note that even if the adversary know that the messages being encrypted is one of two, he still cannot tell the distributions of ciphertext of one message apart from the other.
3. Any cryptosystem in which the encryption algorithm E is deterministic immediately fails to pass this security requirement. (e.g given e, m_0, m_1 and c it would be trivial to decide whether $c = E(e, m_0)$ or $c = E(e, m_1)$ as for each message the ciphertext is unique.)

The next definition is called *Semantic Security.* It may be viewed as a computational version of Shannon's perfect secrecy definition. It requires that the adversary should not gain any computational advantage or partial information from having seen the ciphertext.

Definition 17 *We say that an public key cryptosystem (G, E, D) is semantically secure if $\forall$ PPT algorithm A $\exists$ PPT algorithm B, s.t. $\forall$ PPT algorithm M, $\forall$ function $h : M(1^k) \to \{0,1\}^*$, $\forall c > 0$, $\exists k_0$, $\forall k > k_0$, $\Pr[A(e, |m|, c) = h(m) \mid (e,d) \in G(1^k);\, m \in M(1^k);\, c \in E(e,m)] \leq \Pr[B(e, |m|) = h(m) \mid m \in M(1^k)] + \frac{1}{k^c}$.*

The algorithm M corresponds to the message space from which messages are drawn, and the function $h(m)$ corresponds to information about message m (for example, $h(m) = 1$ if m has the letter 'e' in it).

Theorem 18 [31, 67] *A Public Key Cryptosystem is computationally indistinguishable if and only if it is semantically secure.*

5.2. Encryption: construction phase

We turn now to showing how to actually build a public key encryption scheme which is polynomial time indistinguishable. The construction shown here is by Goldwasser and Micali [31]. The key to the construction is to answer a simpler problem: how to securely encrypt single bits. Encrypting general messages would follow by viewing each message as a string of bits each encrypted independently.

Given a collection of trapdoor predicates B, we define a public key cryptosystem $(G, E, D)_B$ as follows:

Definition 19 *A probabilistic encryption $PE_B = (G, E, D)$ based on trapdoor predicates B is defined as:*

1. *Key generation algorithm G: On input 1^k, G outputs (i, t_i) where $B_i \in B$, $i \in \{0,1\}^k$ and t_i is the trapdoor information. The public encryption key is i and the private decryption key is t_i. (This is achieved by running the sampling algorithm S_1 from the def of B.)*
2. *Let $m = m_1 \dots m_n$ where $m_j \in \{0,1\}$ be the message.*
   $E(i, m)$ encrypts m as follows:
   Choose $x_j \in_R D_i$ such that $B_i(x_j) = m_j$ for $j = 1, \dots, n$.
   Output $c = f_i(x_1) \dots f_i(x_n)$.
3. *Let $c = y_1 \dots y_k$ where $y_i \in D_i$ be the cyph ertext.*
   $D(t_i, c)$ decrypts c as follows:
   Compute $m_j = B_i(y_j)$ for $j = 1, \dots, n$.
   Output $m = m_1 \dots m_n$.

It is clear that all of the above operations can be done in expected polynomial time from the definition of trapdoor predicates and that messages can indeed be sent this way.

Let us ignore for a minute the apparent inefficiency of this proposal in bandwidth expansion and computation (which has been addressed by Blum and Goldwasser in [11]) and talk about security. It follows essentially verbatim from the definition of trapdoor predicates that this system is polynomially time indistinguishable in the case the message is a single bit (i.e. $n = 1$). Even though every bit individually is secure, it is possible in principle that some predicate computed on all the bits (e.g. their parity) is easily computable. Luckily, it is not the case.

We prove polynomial time indistinguishability using the *hybrid argument.* This method is a key proof technique in the theory of pseudo randomness and secure protocol design, in enabling to show how to convert a slight "edge" in solving a problem into a complete surrender of the problem.

As this is one of the most straight forward simplest examples of this technique we shall give it in full.

Theorem 20 [31] *Probabilistic encryption $PE_B = (G, E, D)$ is semantically secure if and only if B is a collection of trapdoor predicates.*

Proof Suppose that (G, E, D) is not indistinguishably secure (i.e. not semantically secure). Then there is a $c > 0$, a PPT A and M such that for infinitely many k, $\exists m_0, m_1 \in M(1^k)$ with $|m_0| = |m_1|$,

$$(*) \qquad \Pr[A(i,c) = 1 \text{ where } (i,t_i) \in G(1^k);\ c \in E(i,m_0)]$$
$$- \Pr[A(i,c) = 1 (i,t_i) \in G(1^k);\ c \in E(i,m_1)] \geq \frac{1}{k^c},$$

where the probability is taken the choice of (i,t_i), the coin tosses of A and E.

Consider k where (*) holds. Wlog, assume that $|m_0| = |m_1| = k$ and that A says 0 more often when c is an encryption of m_0 and 1 more often when c is an encryption of m_1.

Define distributions $D_j = E(i, s_j)$ for $j = 0, 1, \ldots, k$ where $s_0 = m_0, s_k = m_1$ and s_j differs from s_{j+1} in precisely 1 bit.

Let $P_j = \Pr[A(i,c) = 1 | c \in D_j]$.

Then $P_k - P_0 \geq \frac{1}{k^c}$ and since $\sum_{j=0}^{k-1}(P_{j+1} - P_j) = P_k - P_0$, $\exists j$ such that $P_{j+1} - P_j \geq \frac{1}{k^{c+1}}$.

Assume that s_j and s_{j+1} differ in the l^{th} bit; that is, $s_{j,l} \neq s_{j+1,l}$ or, equivalently, $s_{j+1,l} = \bar{s_{j,l}}$ where $s_{j,u}$ is the u-th bit of s_j.

Now, consider the following algorithm B which takes input i, y and outputs 0 or 1 as its guess to the value of the hard core predicate $B_i(y)$.

B on input i, y:

1. Choose $y_1, \ldots, y_k$ such that $B_i(y_r) = s_{j,r}$ for $r = 1, \ldots, k$ using S_1 from the definition of B.
2. Let $c = y_1, \ldots, y, \ldots, y_k$ where y has replaced y_l in the l^{th} block.
3. If $A(1^k, i, , m_0, m_1, c) = 0$ then output $s_{j,l}$.
   If $A(1^k, i, , m_0, m_1, c) = 0$ then output $s_{j+1,l} = \bar{s}_{j,l}$.

Note that $c \in E(i, s_j)$ if $B_i(y) = s_{j,l}$ and $c \in E(i, s_{j+1})$ if $B_i(y) = s_{j+1,l}$.

Thus, in step 3 of algorithm B, outputting $s_{j,l}$ corresponds to A predicting that c is an encryption of s_j.

Claim $\Pr[B(i,y) = B_i(y)] > \frac{1}{2} + \frac{1}{k^{c+1}}$.

Proof

$$\begin{aligned}
\Pr[B(i, f_i(y)) = B_i(y)] &= \Pr[A(i,c) = 0 | c \in E(i, s_j)] \Pr[c \in E(i, s_j)] \\
&\quad + \Pr[A(i,c) = 1 | c \in E(i, s_{j+1})] \Pr[c \in E(i, s_{j+1})] \\
&\geq (1 - P_j)(\frac{1}{2}) + (P_{j+1})(\frac{1}{2}) \\
&= \frac{1}{2} + \frac{1}{2}(P_{j+1} - P_j) \\
&> \frac{1}{2} + \frac{1}{k^{c+1}}.
\end{aligned}$$

Thus, B will predict $B_i(y)$ given i, y with probability better than $\frac{1}{2} + \frac{1}{k^{c+1}}$. This contradicts the assumption that B_i is a trapdoor predicate.

Hence, the probabilistic encryption $PE = (G, E, D)$ is indistinguishably secure.

5.3. Strengthening the adversary: non malleable security

The entire discussion so far has assumed that the adversary can listen to the cipher texts being exchanged over the insecure channel, read the public-file (in the case of public-key cryptography), generate encryptions of any message on his own (for the case of public-key encryption), and perform probabilistic polynomial time computation.

One may imagine a more powerful adversary who can intercept messages being transmitted from sender to receiver and either stop their delivery all together or alter them in some way. Even worse, suppose the adversary can after seeing a ciphertext, request a polynomial number of related ciphertexts to be decrypted for him. For definitions and constructions of encryption schemes secure against such adverdary see [69, 21, 12, 17].

6. A constructive theory of pseudo randomness

A theory of randomness based on computability theory was developed by Kolmogorov, Solomonov and Chaitin [68, 47, 16]. This theory applies to individual strings and defines the complexity of strings as the shortest program (running on a universal machine) that generates that string. A perfectly random string is the extreme case for which no shorter program than the length of the string itself can generate it. Inherintly, it is impossible to generate perfect random strings from shorter ones.

One of the surprising contributions of cryptographically motivated research in the early eighties, has been a theory of randomness computational complexity theory pioneered by Shamir [70] Blum and Micali [10], which makes it possible in principle to deterministically generate random strings from shorter ones. Not to mix notions, we will henceforth refer to this latter development as a theory of pseudo randomness, and the strings generated as pseudo random. In contrast, when we speak of choosing a truly random string of a fixed length over some alphabet, we refer to selecting it with uniform probability over all strings of the same length. In this section we shall only speak of binary alphabet. The notation $x \in_R \{0,1\}^k$ will thus be taken to mean that for every $s \in \{0,1\}^k$, the probability of $x = s$ is $1/2^k$.

Defining pseudo-random distributions is a special case of the definition of computational indistinguishability, which we encountered earlier in the context of secure encryption. A distribution over binary strings is called *pseudo-random* if it is computationally indistinguishable from the uniform distribution over all binary strings of the same length. The idea is that as long as we cannot tell apart samples from the uniform distribution from samples of a distribution X in polynomial time , there is no difference between using either distributions that can be observed in polynomial time. In particular, any probabilistic algorithm, in which the internal coin flips of

the algorithm are replaced by strings sampled from X, must not behave any different than it would using truly random coin flips. A counter example will yield a statistical test to distinguish between X and the uniform distribution.

A deterministic polynomial time program which 'stretchs' a short input string selected with uniform distribution (henceforth called the 'seed'), to a polynomial long output string is called a pseudo random sequence generator. When such a construction is accompanied with a proof that the output string distribution is pseudo random we call the generator a strong pseudo random sequence generator (SPRSG).[5]

In a culmination of a sequence of results by [70, 10, 76, 23, 42], Hastand, Impagliazzo, Levin and Luby showed that a necessary and sufficient condition for the existence of strong pseudo random sequence generators is the existence of one-way functions.

The link between one-way functions and pseudo randomness starts from the following observation. First, rephrase the fact that inverting one-way functions is difficult, by saying that the inverse of a one-way function is unpredictable. In particular, the hard-core of a one-way function is impossible to predict with any non-negligible probability greater than $\frac{1}{2}$. Second, show that impossibility to predict is the ultimate test for pseudo randomness. Namely, if a pseudo-random sequence generator has the property that it is difficult to predict the next bit from previous ones with probability significantly better than $\frac{1}{2}$ in time polynomial in the size of the seed, then it is impossible to distinguish in polynomial time between strings produced by the pseudo random sequence generators and truly random strings. This is proved by turning any statistical test that distinguishes in polynomial time pseudo random strings from random strings into polynomial time next bit predictor. This link is not conditional on the existence of one-way functions. In fact, in work by Nisan and Wigderson [57] they removed the requirement that the pseudo random sequence generator has to work in time which is as fast as the algorithm trying to distinguish the output sequences from truly random. Generators of this type are generally useless for cryptographic applications (as they can not be generated in feasible time) but are very useful for proving complexity theoretic results.

Strong pseudo random generators are useful for understanding the relation between deterministic algorithms and probablistic algorithms. The idea which was put forth by Yao [76] was to replace a single execution of a probablistic polynomial time algorithm A with the majority output of all the executions of the same algorithm, where each execution uses instead of random coins the output of a strong pseudo random number generator on a different input seed. The cost of the latter deterministic procedure will be a factor of $2^{k'}$ longer where k' is the seed length used to generate the pseudo random sequences necessary. The algorithm A must behave "the same" when it uses truly random coins as when it uses coins which are pseudo-random, as otherwise it becomes a distinguisher between the uniform and pseudo-random distributions, an impossible task for a probabilistic polynomial

[5] Again the choice of polynomial-time is arbitrary here, a strong pseudo random sequence generator can be defined to be a deterministic program which works in time $T(n)$ where n is the seed length and is computationally indistinguishable with respect to algorithms which run in time $T'(n)$ for time functions T, T'.

time algorithm. Putting this together, we get : if one-way functions exist, then $BPP \subseteq \cap_\epsilon DTIME(2^{k^\epsilon})$. This tradeoff between the *hardness* of inverting the one-way function, and *randomness* replacement, has been followed up with many papers in complexity theory each either relaxing the hardness assumption or tightening the relation between deterministic and probabilistic complexity classes.

Strong pseudo random generators are particularly useful for cryptography. Suppose you need a large supply of random strings for your cryptographic applications (e.g. the choice of secret keys, internal coin tosses of an encryption algorithm, etc.). If you use instead of truly random bits, pseudo random sequence generators which are weak (e.g. predictable), it may completely destroy the underlying cryptographic applications [14]. In contrast, we can replace any use of truly random coins with strong pseudo random ones (assuming we have access to truly random coins for the seeds — which is an interesting discussion all by itself), without fear of compromising the security of the underlying application. Indeed, if as a result of such replacement the cryptographic application becomes insecure, then a way is found to distinguish outputs of SPRG from the uniform distribution. Many classical pseudo random number generators which are quite useful and effective for Monte Carlo simulations, have been shown not only weak but predictable in a strong sense which makes them typically unsuitable for cryptographic applications. For example, *linear* feedback shift registers [37] are well-known to be cryptographically insecure; one can solve for the feedback pattern given a small number of output bits, and similarily outputs of linear congruential generators [22]. In [44] Kannan, Lenstra, and Lovasz use the L^3 algorithm to show that the binary expansion of any algebraic number y (such as $\sqrt{5} = 10.001111000110111\ldots$) is insecure, since an adversary can identify y exactly from a sufficient number of bits, and then extrapolate y's expansion.

6.1. Pseudo random functions, permutations, and what else?

Similarily to defining pseudo random sequences one may ask what other random objects can be replaced with pseudo-random counter parts. Goldreich, Goldwasser and Micali [23] considered in this light random functions, which from a gold mind for applications. Pseudo random functions are defined to be for every size k a subset of all functions from (and to) the binary strings of length k, which are polynomial time indistinguishable from truly random functions by any algorithm whose only access to the function is to query it on inputs of its choice. However, in contrast with a truly random function, a pseudo random function has a short description which if known enables efficient evaluation.

Let $H_k = \{f : \{0,1\}^k \to \{0,1\}^k\}$ then $|H_k| = (2^k)^{2^k}$. Let $\mathcal{H} = \bigcup_k H_k$.

Definition 21 *A* polynomial time statistical test for functions *is a polynomial time algorithm T^f with access to a black box f from which T can request values of $f(x)$ for x of T's choice. A collection of functions $\mathcal{F} = \bigcup_k F_k$ where $F_k \subset H_k$* passes the statistical test *T if $\forall Q \in \mathbf{Q}[x], \exists k_0, \forall k > k_0\ |T(F_k) - T(H_k)| < \frac{1}{Q(k)}$ where $T(F_k) = \Pr_{f \in F_k, coins\ of\ T}[T^f(1^k) = 1]$ and $T(H_k) = \Pr_{f \in H_k, coins\ of\ T}[T^f(1^k) = 1]$.*

Definition 22 *A collection of functions* $\mathcal{F} = \bigcup_k F_k$ *is a* pseudo-random collection of functions *if*

1. *(Indexing) For each* k*, there is a unique index* $i \in \{0,1\}^k$ *associated with each* $f \in F_k$*. The function* $f \in F_k$ *associated with index* i *will be written* f_i*.*
2. *(Efficiency) There is a polynomial time function* A *so that* $A(i,x) = f_i(x)$*.*
3. *(Pseudo-randomness)* $\mathcal{F}$ *passes all polynomial time statistical tests for functions.*

Theorem 23 [23] *If there exist one-way functions, then there exist pseudo-random collections of functions.*

An immediate application of pseudo random functions is the construction of semantically secure private key cryptosystem as follows. Let s an index of a pseudo random function f_s be the joint secret key of the sender Alice and the receiver Bob. Then to encrypt message m, Alice selects at random $r \in \{0,1\}^k$, and sets the cipher text $c = (r, f_s(r) \oplus m)$ where $\oplus$ is the bit-wise exclusive-or of two strings. To decrypt $c = (a,b)$, Bob computes $f_s(a) \oplus b$.

Pseudo random functions have been used to derive negative results in computational learning theory by Valiant and Kearns [73]. They show that any concept class (i.e. a set of Boolean functions) which contains a family of pseudo random functions cannot be efficiently learnable under the uniform distribution and with the help of membership queries. A learning algorithm is given oracle access to any function in the class and is required to output a description of a function which is close to the target function (being queried).

The work on *natural proofs* originated by Rudich and Razborov [55] use pseudo random functions to derive negative results on the possibility of proving good complexity lower bounds using a restricted class of circuit lower bound proofs referred to as *natural*. It is proved that natural (lower bound) proofs cannot be established for complexity classes containing a family of pseudo random functions.

An interesting question is to characterize which classes of random objects can be replaced by pseudo random objects. Luby and Rackoff [50] treated the case of pseudo random permutations and Naor and Reingold the case of permutations with cyclic structure [58]. As any object can be abstracted as a restricted class of functions, the real question is what form of access to the function does the statistical test have. In the standard definition, the statistical test for functions can query the functions at values of its choice. This may not be necessarily the natural choice in every case. For example, if the function corresponds to the description of a random graph (e.g. $f(u,v) = 1$ if and only if an edge is present between vertices u and v).

Define the "ultimate" extension of a statistical test for functions on k bit strings, to be given access to the *entire truth table* of the function (i.e. an exponential size input). The following observation is then straightforward.

Theorem 24 *Let* $f : \{0,1\}^* \to \{0,1\}^*$ *be polynomial time computable function, for which the fastest inverting algorithm runs in time* 2^{n^ϵ} *for some* $\epsilon > 0$*. Then, there exist collections of pseudo random functions which pass all ultimate statistical tests for functions.*

7. Interactive protocols, interactive proofs, and zero knowledge interactive proofs

Secure one-way communication is a special case of general interactive protocols. The most exciting developments in cryptography beyond public-key cryptography has been the development of interactive protocols, interactive proofs, and zero knowledge interactive proofs [32, 38, 76, 35, 8, 19, 13]. [6] Unfortunately, we have no space to cover these developments in this article. These topics have been surveyed extensively, and the interested reader may turn to [39, 40].

A few final words. Generally speaking, an *interactive protocol* consists of two or more parties who cooperate and coordinate without a trusted "third" party to accomplish a common goal, referred to as the *functionality* of the protocol, while maintaining the *secrecy* of their private data. A functionality may be computing a simple deterministic function such as majority of the inputs of the communicating parties, or a more complicated probabilistic computation such as playing a non-cooperative game without a trusted referee.

In the case of more than two parties, the case of adversarial coalitions of participants who attempt to damage the functionality and break secrecy has been considered. Very powerful and surprising theorems about the ability of playing non-cooperative games without a trusted "third party" have been shown. A sample theorem of Benor, Goldwasser, and Wigderson shows that in the presence of an adversarial coalition containing less than a third of the parties, any probabilistic computation can be performed maintaining functionality and perfect information theoretic secrecy of the inputs, as long as each pair of parties can communicate in perfect secrecy [8, 19]. These results make extensive use of error correcting codes based on polynomials. The connection between these theorems and research in game theory and threory of auctions is well worth examining.

References

[1] W. B. Alexi, B. Chor, O. Goldreich, and C. P. Schnorr. RSA/Rabin functions: certain parts are as hard as the whole. *SIAM J. Computing*, 17(2):194–209, April 1988.

[2] L. M. Adleman. Algorithmic Number Theory — The Complexity Contribution Proceedings of the Foundations of Computer Science, 88–99, October 1994.

[3] L. M. Adleman and M. A. Huang. Recognizing primes in random polynomial time. In *Proc. 19th ACM Symp. on Theory of Computing*, 462–469, New York City, 1987. ACM.

[4] M. Ajtai. Generating Hard Instances of Lattice Problems. In *28th STOC*, 99–108, 1996.

[5] D. Angluin. Lecture notes on the complexity of some problems in number the-

[6] In particular, the idea of multi-prover interactive proofs of Benor, Goldwasser, Kilian, and Wigderson [15] (which has become better known as *probabilistic checkable proofs*) has led to a rice body of NP-hardness results for approximation versions of optimization problems [43].

ory. Technical Report TR-243, Yale University Computer Science Department, August 1982.

[6] D. Angluin, D. Lichtenstein, Provable security of cryptosystems:A survey. Tech. Rep. 288, Dept. of Computer Science, Yale Univ. New Haven, Conn., 1983.

[7] Eric Bach. How to generate factored random numbers. *SIAM J. Computing*, 17(2):179–193, April 1988.

[8] M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for fault-tolerant distributed computing. In *Proc. 20th ACM Symp. on Theory of Computing*, 1–10, Chicago, 1988. ACM.

[9] M. Blum. Coin flipping by telephone. In *Proc. IEEE Spring COMPCOM*, 133–137. IEEE, 1982.

[10] M. Blum and S. Micali. How to generate cryptographically strong sequences of pseudo-random bits. *SIAM J. Computing*, 13(4):850–863, November 1984.

[11] M. Blum and S. Goldwasser. An Efficient Probabilistic Public-Key Encryption Scheme which hides all partial information. In *Crypto84*, LNCS (196) Springer-Verlag, 289–302.

[12] M. Blum and P. Feldman and S. Micali, Proving Security Against Chosen Cyphertext Attacks, In proceedings of CRYPTO88, 256–268, 1988.

[13] L. Babai. Trading Group Theory for Randomness. In *17th STOC*, 421–420, 1985.

[14] Bellare, M. and Goldwasser, S. and Micciancio, D., "Pseudo-Random" Number Generation within Cryptographic A lgorithms: The DSS Case, In proceedings of Crypto '97, 277–291, 1977.

[15] M. Ben-Or, S. Goldwasser,J. Kilian, and A. Wigderson, Multi-Prover Interactive Proof-Systems, Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing (1988), 113–131.

[16] C.J. Chaitin On the Length of Programs for Computing Finite Binary Sequences. Journal of the ACM, vol 13, 547–570, 1966.

[17] R. Cramer, V. Shoup, "A Practical Public Key Cryptosystem Provably Secure against Adaptive Chosen Ciphertext Attack", Advances in Cryptology — CRYPTO '98 Proceedings, 13–25, Springer-Verlag, 1998

[18] Christian Cachin and Ueli Maurer Unconditional Security Against Memory-Bounded Adversaries Advances in Cryptology — CRYPTO '97, Lecture Notes in Computer Science, Springer-Verlag, vol. 1294, 292–306, 1997.

[19] Crepeau D. Chaum and I. Damgard. Multiparty unconditionally secure protocols. In Proc. of 20th ACM Symp. on Theory of Computing, Chicago, 1988.

[20] W. Diffie and M. E. Hellman. New directions in cryptography. *IEEE Trans. Inform. Theory*, IT-22:644–654, November 1976.

[21] D. Dolev, C. Dwork, and M. Naor. Non-malleable cryptography. In *Proc. 23rd ACM Symp. on Theory of Computing*, 542–552. ACM, 1991.

[22] A. M. Frieze, J. Hastad, R. Kannan, J. C. Lagarias, and A. Shamir. Reconstructing truncated integer variables satisfying linear congruences. *SIAM J. Computing*, 17(2):262–280, April 1988.

[23] O. Goldreich, S. Goldwasser, and S. Micali. How to Construct Random Func-

tions. *JACM*, Vol. 33, No. 4, 792–807, 1986.
[24] O. Goldreich. *Foundations of Cryptography: Volume 1 – Basic Tools.* Cambridge University Press, 2001.
[25] Oded Goldreich. newblock A Note on Computational Indistinguishability. Information Processing Letters 34(6): 277–281 (1990)
[26] Oded Goldreich, Russell Impagliazzo, Leonid A. Levin, Ramarathnam Venkatesan, David Zuckerman. Security Preserving Amplification of Hardness. FOCS 1990: 318–326
[27] Oded Goldreich, Hugo Krawczyk. Sparse Pseudorandom Distributions. CRYPTO 1989: 113–127
[28] O. Goldreich and L. Levin. A hard-core predicate for all one-way functions. Proc., ACM Symp. on Theory of Computing, 25–32, 1989.
[29] S. Goldwasser, S. Micali, and Ronald L. Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM J. Computing*, 17(2):281–308, April 1988.
[30] S. Goldwasser and J. Kilian. Almost all primes can be quickly certified. In *Proc. 18th ACM Symp. on Theory of Computing*, 316–329, Berkeley, 1986. ACM.
[31] S. Goldwasser and S. Micali. Probabilistic encryption. *JCSS*, 28(2):270–299, April 1984.
[32] S. Goldwasser, S. Micali, and C. Rackoff. The knowledge complexity of interactive proof-systems. *SIAM. J. Computing*, 18(1):186–208, February 1989.
[33] S. Goldwasser, and Y. Lindell. Proceedings of 16th International Symposium on Distributed Computing to appear October 2002.
[34] Goldwasser, Goldreich, Halevi. Collision-Free Hashing from Lattice Problems. Electronic Colloquium on Computational Complexity (ECCC) 3(42): (1996)
[35] O. Goldreich, S. Micali and A. Wigderson. How to Play any Mental Game — A Completeness Theorem for Protocols with Honest Majority. In *19th STOC,* 218–229, 1987.
[36] S. Goldwasser, S. Micali, and P. Tong. Why and how to establish a private code on a public network. In *Proc. 23rd IEEE Symp. on Foundations of Comp. Science*, 134–144, Chicago, 1982. IEEE.
[37] S. W. Golomb. *Shift Register Sequences.* Aegean Park Press, Laguna Hills, 1982. Revised edition.
[38] O. Goldreich, S. Micali, and A. Wigderson. Proofs that yield nothing but their validity and a methodology of cryptographic protocol design. In *Proc. 27th IEEE Symp. on Foundations of Comp. Science*, 174–187, Toronto, 1986. IEEE.
[39] O. Goldreich, The Foundations of Cryptography - Volume 1 ISBN 0-521-79172-3 Cambridge University Press.
[40] O. Goldreich Modern Cryptography, Probabilistic Proofs and Pseudorandomness ISBN 3-540-64766-x Springer-Verlag, Algorithms and Combinatorics, Vol 17, 1998.
[41] Y. Gertner, T. Malkin and O. Reingold. On the impossibility of basing trapdoor functions on trapdoor predicates, FOCS 2001.
[42] J. Hastad, R. Impagliazzo, L. Levin, Michael Luby. Construction of a pseudo-

random generator from any one-way function. SIAM J. Comput. 28(4):1364–1396, 1999.
[43] D. Johnson Tale of the Second Prover. Journal of Algorithms. Vol 13.
[44] R. Kannan, A. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. In *Proc. 16th ACM Symp. on Theory of Computing*, 191–200, Washington, D.C., 1984. ACM.
[45] J. Kilian. Founding cryptography on oblivious transfer. In *Proc. 20th ACM Symp. on Theory of Computing.*
[46] N. Koblitz, A. Menezes, and S. Vanstone. The state of elliptic curve cryptography. Designs, Codes and Cryptography, 19 (2000), 173–193.
[47] A. Kolmogorov Three approaches to the concept of the amount of information Probl. of Inform. Trandm., Vol1/1/,1965.
[48] A. K. Lenstra and H. W. Lenstra, Jr. Algorithms in number theory. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science (Volume A: Algorithms and Complexity)*, chapter 12, 673–715. Elsevier and MIT Press, 1990.
[49] A.K. Lenstra, H.W. Lenstra, L. Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen* 261, 515–534 (1982).
[50] M. Luby and C. Rackoff, Pseudo-Random Permutation Generators and Cryptographic Composition. In proceedings of STOC86, 356–363, 1986.
[51] Ueli Maurer, Protocols for Secret Key Agreement by Public Discussion Based on Common Information IEEE Trans. on Inform. Theory (1993).
[52] Naor and Yung, Universal One-Way Hash Functions and their Cryptographic Applications Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing. (May 15–17 1989: Seattle, WA, USA)
[53] Noam Nisan. Pseudorandom generators for space-bounded computation. Proceedings of the Twenty Second Annual ACM Symposium on Theory of Computing, 204–212, Baltimore, Maryland, 14–16 May 1990.
[54] A. M. Odlyzko, Discrete logarithms: The past and the future, Designs, Codes, and Cryptography 19, 129–145, 2000
[55] A.R. Razborov and S. Rudich. Natural proofs. Journal of Computer and System Science, Vol. 55 (1), 24–35, 1997.
[56] Ronald L. Rivest, Adi Shamir, and Leonard M. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.
[57] N. Nisan and A. Wigderson, Hardness vs. Randomness Journal of JCSS, Vol 49, No 2, 149–167, 1994.
[58] M. Naor and O. Reingold, Constructing pseudorandom permutations with a prescribed structure, SODA 2001, 458–459, 2001.
[59] A. K. Lenstra, H. W. Lenstra, Jr., M. S. Manasse, and J. M. Pollard. The number field sieve. In *Proc. 22nd ACM Symp. on Theory of Computing*, 564–572, Baltimore, Maryland, 1990. ACM.
[60] John Rompel. One-way functions are necessary and sufficient for secure signatures. Proceedings of the Twenty Second Annual ACM Symposium on Theory

of Computing, 387–394, Baltimore, Maryland, 14–16 May 1990.
[61] T. Rabin and M. Ben-Or. Verifiable secret sharing and multiparty protocols with honest majority. In *21st ACM Symposium on Theory of Computing*, 73–85, 1989.
[62] M. Rabin. Digitalized signatures as intractable as factorization. Technical Report MIT/LCS/TR-212, MIT Laboratory for Computer Science, January 1979.
[63] Madhu Sudan. Coding Theory: Tutorial and Survey. Proceedings of the 42nd Annual Symposium on Foundations of Computer Science, 36–53, Las Vegas, Nevada, 14–17 October, 2001.
[64] C.P. Schnorr. A hierarchy of polynomial time lattice basis reduction algorithms. In *Theoretical Computer Science*, vol. 53, 1987, 201–224.
[65] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:623–656, 1948.
[66] C. E. Shannon. Communication theory of secrecy systems. *Bell Sys. Tech. J.*, 28:657–715, 1949.
[67] S. Micali, C. Rackoff, and R. H. Sloan. The notion of security for probabilistic cryptosystems. *SIAM J. Computing*, 17(2):412–426, April 1988.
[68] R.J. Solomonoff A formal theory of Inductive Inference. Inform. and Control. Vol 7/1, 1–22, 1964.
[69] Rackoff, C. and Simon, D. R., Non-interactive zero-knowledge proof of knowledge and chosen ciphertext attack, YEAR = 1991, Proceedings of Crypto '91, 433–444,
[70] A. Shamir. On the generation of cryptographically strong pseudo-random sequences. In *Proc. ICALP*, 544–550. Springer, 1981.
[71] A. Sahai and S. Vadhan. A Complete Promise Problem for Statistical Zero-Knowledge. In *38th FOCS*, 448–457, 1997.
[72] Luca Trevisan Extractors and Pseudorandom Generators J. of the ACM, 48(4):860–879, 2001.
[73] M. Kearns, and L. Valiant. Cryptographic limitations on learning Boolean formulae and finite automata, J. Assoc. Comp. Mach., 41:1 (1994) 67–95.
[74] H. Williams, "A Modification of the RSA Public-Key Encryption Procedure", IEEE Trans. Information Theory, 26(6) (1980), 726–729.
[75] A. D. Wyner, The wire-tap channel. Bell System Technical Journal, Vol. 54, no. 8, 1975, 1355–1387.
[76] A. C. Yao. Theory and application of trapdoor functions. In *Proc. 23rd IEEE Symp. on Foundations of Comp. Science*, 80–91, Chicago, 1982. IEEE.

ICM 2002 · Vol. I · 273–290

Random Matrices, Free Probability and the Invariant Subspace Problem Relative to a von Neumann Algebra

U. Haagerup*

2000 Mathematics Subject Classification: 46L35, 46L54, 46L80, 47A15, 47C15, 60B99, 81S30.
Keywords and Phrases: C^*-algebras, von Neumann algebras, Random matrices, Free probability, Invariant subspaces.

1. Introduction

Random matrices have their roots in multivariate analysis in statistics, and since Wigner's pioneering work [Wi] in 1955, they have been a very important tool in mathematical physics. In functional analysis, random matrices and random structures have in the last two decades been used to construct Banach spaces with surprising properties. After Voiculescu in 1990–1991 used random matrices to classification problems for von Neumann algebras, they have played a key role in von Neumann algebra theory (cf. [V8]). In this lecture we will discuss some new applications of random matrices to operator algebra theory, namely applications to classification problems for C^*-algebras and to the invariant subspace problem relative to a von Neumann algebra.

The rest of this lecture is divided into eight sections:

2. Selfadjoint random matrices and Wigner's semicircle law.
3. Free probability and Voiculescu's random matrix model.
4. $\mathrm{Ext}(C_r^*(F_k))$ is not a group for $k \geq 2$.
5. Other applications of random matrices to C^*-algebras.
6. The invariant subspace problem relative to a von Neumann algebra.
7. The Fuglede-Kadison determinant and Brown's spectral distribution measure.
8. Spectral subspaces for operators in II_1-factors.
9. Voiculescu's circular operator Y and the strictly upper triangular operator T.

*Department of Mathematics & Computer Science, University of Southern Denmark, Campusvej 55, DK–5230 Odense M, Denmark. E-mail: haagerup@imada.sdu.dk

2. Selfadjoint random matrices and Wigner's semicircle law

A random matrix X is an $n \times n$ matrix whose entries are real or complex random variables on a probability space $(\Omega, \mathcal{F}, P)$. We denote by $\mathrm{SGRM}(n, \sigma^2)$ the class of selfadjoint random matrices

$$X_n = (X_{ij}^{(n)})_{i,j=1}^n$$

where X_{ij}, $i, j = 1, \ldots, n$ are n^2 complex random variables and

$$(X_{ii}^{(n)})_i, \quad (\sqrt{2}\,\mathrm{Re}\, X_{ij}^{(n)})_{i<j}, \quad (\sqrt{2}\,\mathrm{Im}\, X_{ij}^{(n)})_{i<j}$$

are n^2 independent identical distributed real Gaussian random variables with mean value 0 and variance σ^2. In the terminology of Mehta's book [Me], X_n is a Gaussian unitary ensemble (GUE). In the following we put $\sigma^2 = \frac{1}{n}$ which is the normalization used in Voiculescu's random matrix paper [V4]. By results of Gaudin, Mehta and Wigner from 1960–1965, the joint distribution of the eigenvalues (in random order) of X has density g given by

$$g_n(\lambda_1, \ldots, \lambda_n) = c_n \prod_{i<j} (\lambda_j - \lambda_i)^2 \exp\Big(-\frac{n}{2} \sum_{i=1}^n \lambda_i^2\Big)$$

where c_n is a normalization constant, and the (average) density for a single eigenvalue is given by

$$h_n(x) = \frac{1}{\sqrt{2n}} \sum_{k=0}^{n-1} \varphi_k\Big(\sqrt{\frac{n}{2}}\, x\Big)^2$$

where $\varphi_0, \varphi_1, \ldots$ is the sequence of Hermite functions. Moreover,

$$\lim_{n\to\infty} h_n(x) = \frac{1}{2\pi} \sqrt{4 - x^2}\, 1_{[-2,2]}(x), \quad x \in \mathbb{R}$$

(cf. [Me]). This is Wigner's semicircle law for the GUE-case. In the sense of weak convergence of probability measures, the semicircle law can be proved under much more general assumptions on the entries (see Wigner [Wi]). Arnold proved in 1967 that the corresponding strong law also holds, i.e. for almost all ω in the probability space Ω, the empirical eigenvalue distribution of $X_n(\omega)$ converges weakly to the semicircular distribution $\frac{1}{2\pi}\sqrt{4 - x^2}\, 1_{[-2,2]}(x)dx$ as $n \to \infty$. Very interesting research have been carried out on the level spacing of the eigenvalues in the bulk of the spectrum (cf. [Me]) and more recently near the boundary of the spectrum (cf. [TW1], [TW2]) for selfadjoint Gaussian random matrices with real, complex or symplectic entries (the GOE, GUE and GSE cases), but this is outside the scope of the present lecture.

3. Free probability and Voiculescu's random matrix model

Voiculescu proved in 1991 [V4] an extensive generalization of Wigner's semicircle law to families of independent random matrices. In order to state the result, we will need some basic concepts from free probability theory (cf. [V2], [V3] and [VDN]).

Definition 3.1 [V2]

1. *A non-commutative probability space is a pair* (A, φ) *consisting of a unital complex algebra* A *and a functional* $\varphi\colon A \to \mathbb{C}$ *such that* $\varphi(1_A) = 1$.
2. *A* C^**-probability space is a pair* (A, φ) *consisting of a unital* C^**-algebra* A *and a state* $\varphi\colon A \to \mathbb{C}$ *on* A.

The connection to classical probability theory on a probability space $(\Omega, \mathcal{F}, P)$ is obtained by putting

$$A = \bigcap_{p=1}^{\infty} L^p(\Omega)$$

and

$$\varphi(a) = \mathbb{E}(a) = \int_\Omega a(\omega) dP(\omega), \quad a \in A$$

or $A' = L^\infty(\Omega, P)$ with the same definition of φ. The latter example is a C^*-probability space. To fit random matrices (of size n) into this framework, one must instead consider the non-commutative algebra

$$A_n = \bigcap_{p=1}^{\infty} L^p(\Omega, M_n(\mathbb{C}))$$

with functional

$$\varphi_n(a) = \mathbb{E}(\mathrm{tr}_n(a)) = \int_\Omega \mathrm{tr}_n(a(\omega)) d\omega$$

where $\mathrm{tr}_n = \frac{1}{n}\mathrm{Tr}$ is the normalized trace on $M_n(\mathbb{C})$.

Definition 3.2 [V2], [V3]

1. *A family* $(a_i)_{i\in 1}$ *of elements in a non-commutative probability space is a* free family *if for all* $n \in \mathbb{N}$ *and all polynomials* $p_1, \ldots, p_n \in \mathbb{C}[X]$, *one has*

$$\varphi(p_1(a_{i_1}) \cdot \ldots \cdot p_n(a_{i_n})) = 0$$

whenever $i_1 \neq i_2 \neq \cdots \neq i_n$ *(neighbouring indices are different) and* $\varphi(p_k(a_{i_k})) = 0$ *for* $k = 1, \ldots, n$.
2. *A family* $(x_i)_{i\in j}$ *of elements in a* C^**-probability space* (A, φ) *is called a semicircular family if* $(x_i)_{i\in I}$ *is a free family,* $x_i = x_i^*$, $\varphi(x_i^{2k-1}) = 0$ *and*

$$\varphi(x_i^{2k}) = \frac{1}{2\pi} \int_{-2}^{2} t^k \sqrt{4 - t^2} dt = \frac{1}{k+1} \binom{2k}{k}$$

for all $k \in \mathbb{N}$ *and all* $i \in I$.

We can now formulate Voiculescu's generalization of Wigner's semicircle law:

Theorem 3.3 [V4] *Let I be an index set and let for each $n \in \mathbb{N}$, $(X_i^{(n)})_{i\in I}$ be a family of independent SGRM$(n, \frac{1}{n})$-distributed selfadjoint random matrices. Then asymptotically as $n \to \infty$ $(X_i^{(n)})_{i\in I}$ is a semicircular family, i.e. if $(x_i)_{i\in I}$ is a semicircular family index by I in a C^*-probability space (A, φ) then*

$$\lim_{n\to\infty} \mathbb{E}\mathrm{tr}_n(X_{i_1}^{(n)} \cdot \ldots \cdot X_{i_p}^{(n)}) = \varphi(x_{i_1} \cdot \ldots \cdot x_{i_p}) \tag{3.1}$$

for all $p \in \mathbb{N}$ and all $i_1, \ldots, i_p \in I$.

The corresponding strong law: For almost all $\omega \in \Omega$, one has

$$\lim_{n\to\infty} \mathrm{tr}_n(X_{i_1}^{(n)}(\omega) \cdot \ldots \cdot X_{i_p}^{(n)}(\omega)) = \varphi(x_{i_1} \cdot \ldots \cdot x_{i_p}), \tag{3.2}$$

whick was proved independently by Hiai and Petz [HP2] and Thorbjrnsen [T].

4. $\mathrm{Ext}(C_r^*(F_k))$ is not a group for $k \geq 2$

Very recently Thorbjrnsen and the lecturer proved that the strong version (3.2) of Voiculescu's random matrix model also holds for the operator norm:

Theorem 4.1 [HT4] *Let $r \in \mathbb{N}$ and let for each $n \in \mathbb{N}$ $(X_1^{(n)}, \ldots, X_r^{(n)})$ be a set of r independent SGRM$(n, \frac{1}{n})$-distributed selfadjoint random matrices. Let further $(x_1, \ldots, x_r)$ be a semicircular system in a C^*-probability space (A, φ), where φ is a faithful state on A. Then there is a null set $N \subseteq \Omega$ such that for all $\omega \in \Omega\backslash N$ and all non-commutative polynomials P in r variables*

$$\lim_{n\to\infty} \|P(X_1^{(n)}(\omega), \ldots, X_r^{(n)}(\omega))\| = \|P(x_1, \ldots, x_r)\|.$$

Let Γ be a countable (discrete) group. The reduced group C^*-algebra $C_r^*(\Gamma)$ is the C^*-subalgebra of $B(\ell^2(\Gamma))$ generated by the set of unitaries $\{\lambda(\gamma) \mid \gamma \in \Gamma\}$, where $\lambda\colon \Gamma \to B(\ell^2(\Gamma))$ is the left regular representation. By the methods of [V3] it follows that for the free group F_k on k generators, $C_r^*(F_k)$ can be embedded in $C^*(x_1, \ldots, x_k, 1)$, where $x_1, \ldots, x_k$ is a free semicircular family in a C^*-probability space (A, φ) with φ faithful. Hence as a corollary of Theorem 4.1 we have

Corollary 4.2 [HT4] *czj Let $k \in \mathbb{N}$, $k \geq 2$. Then $C_r^*(F_k)$ can be embedded in the quotient C^*-algebra $\prod M_n(\mathbb{C})/\sum M_n(\mathbb{C})$ where*

$$\begin{aligned} \prod M_n(\mathbb{C}) &= \left\{(x_n)_{n=1}^\infty \mid x_n \in M_n(\mathbb{C}),\ \sup_n \|x_n\| < \infty\right\} \\ \sum M_n(\mathbb{C}) &= \left\{(x_n)_{n=1}^\infty \mid x_n \in M_n(\mathbb{C}),\ \lim_{n\to\infty} \|x_n\| = 0\right\}. \end{aligned}$$

In particular $C_r^*(F_k)$ is a MF-algebra in the sense of Blackadar and Kirchberg [BK].

The invariant $\mathrm{Ext}(A)$ for a C^*-algebra A was introduced by Brown, Douglas and Fillmore in [BDF]. $\mathrm{Ext}(A)$ is the set of all essential extensions B of A by the compact operators K on the Hilbert space $\ell^2(\mathbb{N})$, and it has a natural semigroup structure. Voiculescu proved in [V1] that $\mathrm{Ext}(A)$ is always a unital semigroup, and by Choi and Effros [CE] $\mathrm{Ext}(A)$ is a group, when A is a nuclear C^*-algebra. Andersen [An] provided in 1978 the first example of a C^*-algebra A for which $\mathrm{Ext}(A)$ is not a group. The C^*-algebra in [An] is generated by $C_r^*(F_2)$ and a projection $p \in B(\ell^2(F_2))$. Since then it has been an open problem whether $\mathrm{Ext}(C_r^*(F_2))$ is a group (see [V6, Sect.5] for a more detailed discussion about this problem). It is well known that a proof of Corollary 4.2 would provide a negative solution to this problem (see [V6, 5.12], [V5] and [Ro]). The argument works for all $k \geq 2$. Hence we have

Corollary 4.3 [HT4] *For all $k \in \mathbb{N}$, $k \geq 2$, $\mathrm{Ext}(C_r^*(F_k))$ is not a group.*

Remarks 4.4
a) Corollaries 4.2 and 4.3 also hold for $k = \infty$.
b) $C_r^*(F_k)$ is not quasidiagonal (cf [Ro]) but the non-invertible extension B of $C_r^*(F_k)$ obtained from Corollary 4.2 is quasidiagonal.
c) $C_r^*(F_k)$ is an exact C^*-algebra, but for any non-invertible extension B of $C_r^*(F_k)$ by the compact operators, B cannot be exact. This follows from the Lifting theorem in [EH]. Other examples of non-exact extensions of exact C^*-algebras by K are given in [Ki2].

In the rest of this section, I will briefly outline the main steps in the proof of Theorem 4.1. From (3.2) it follows that for all non-commutative polynomials P in r variables

$$\liminf_{n\to\infty} \|P(X_1^{(n)}(\omega), \ldots, X_r^{(n)}(\omega))\| \geq \|P(x_1, \ldots, x_r)\| \tag{4.1}$$

for almost all $\omega \in \Omega$ (see [T]), so we "only" have to prove that

$$\limsup_{n\to\infty} \|P(X_1^{(n)}(\omega), \ldots, X_r^{(n)}(\omega)\| \leq \|P(x_1, \ldots, x_r)\| \tag{4.2}$$

for almost all $\omega \in \Omega$. Even the case $r = 1$ and $P(x) = x$ is a difficult task. It corresponds to proving that if X_n is $\mathrm{SGRM}(n, \frac{1}{n})$-distributed, $n = 1, 2, \ldots$ then for almost all $\omega \in \Omega$,

$$\limsup_{n\to\infty} \lambda_{\max}(X_n(\omega)) \leq 2 \qquad \liminf_{n\to\infty} \lambda_{\min}(X_n(\omega)) \geq -2,$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the smallest and largest eigenvalue of $X_n(\omega)$. This problem was settled by Bai and Yin [BY] in 1988 using Geman's combinatorial method [Ge]. (See also [Ba, Thm. 2.12] and [HT1, Thm. 3.1]).

Lemma 4.5 (The linearization trick) [HT4] *In order to prove (4.2) it is sufficient to show that for all $m \in \mathbb{N}$ and all selfadjoint $m \times m$-matrices $a_0, \ldots, a_r$ and all $\varepsilon > 0$,*

$$\sigma\big(a_0 \otimes 1 + \sum_{i=1}^{r} a_i \otimes X_i^{(n)}(\omega)\big) \subseteq \sigma\big(a_0 \otimes 1 + \sum_{i=1}^{r} a_i \otimes x_i\big) +]-\varepsilon, \varepsilon[\tag{4.3}$$

holds eventually as $n \to \infty$ for almost all $\omega \in \Omega$. Here $\sigma(T)$ denotes the spectrum of a matrix or an operator T.

Lemma 4.6 [HT4] *Let $a_0, \ldots, a_r$ be as above, and put*

$$\begin{aligned} S_n &= a_0 \otimes 1 + \sum_{i=1}^{r} a_i \otimes X_i^{(n)} \\ s &= a_0 \otimes 1 + \sum_{i=1}^{r} a_i \otimes x_i. \end{aligned}$$

Moreover, let G_n, G be the matrix valued Stieltjes transforms of S_n and S, i.e. for $\lambda \in M_n(\mathbb{C})$, and $\operatorname{Im}\lambda = \frac{1}{2i}(\lambda - \lambda^*)$ positive definite

$$\begin{aligned} G_n(\lambda) &= \mathbb{E}((id_m \otimes \operatorname{tr}_n)((\lambda \otimes 1 - S_n)^{-1})) \\ G(\lambda) &= (id_m \otimes \varphi)((\lambda \otimes 1 - s)^{-1}). \end{aligned}$$

Then $G_n(\lambda)$ and $G(\lambda)$ are invertible and

$$a_0 + \sum_{i=1}^{r} a_i G(\lambda) a_i + G(\lambda)^{-1} = \lambda \tag{4.4}$$

$$\left\| a_0 + \sum a_i G_n(\lambda) a_i + G_n(\lambda)^{-1} - \lambda \right\| \leq \frac{C}{n^2}(K + \|\lambda\|)^2 \|(Im\lambda)^{-1}\|^5 \tag{4.5}$$

where $C = \frac{\pi^2 m^3}{8}\left(\sum_{i=1}^{r} \|a_i\|^2\right)^2$ and $K = \|a_0\| + 4\sum_{i=1}^{r} \|a_i\|$.

The equality (4.4) was proved by Lehner (cf. [Le, Prop.4.1] using Voiculescu's R-transform with amalgamation [V7]. The inequality (4.5) is more difficult. It relies on the concentration phenomena used in Banach space theory, in form of [P1, Theorem 4.7]. (See [Mi] for a general discussion of the concentration phenomena.) Next we derive from (4.4) and (4.5) that

$$\|G_n(\lambda) - G(\lambda)\| \leq \frac{4C}{n^2}(K + \|\lambda\|)^2 \|(\operatorname{Im}\lambda)^{-1}\|^7 \tag{4.6}$$

when $\lambda \in M_m(\mathbb{C})$ and $\operatorname{Im}\lambda$ is positive definite. The estimate (4.6) implies that for every $f \in C_c^\infty(\mathbb{R})$

$$\mathbb{E}((\operatorname{tr}_m \otimes \operatorname{tr}_n)(f(S_n))) = (\operatorname{tr}_m \otimes \varphi)(f(s)) + O\left(\frac{1}{n^2}\right) \tag{4.7}$$

for $n \to \infty$. Moreover a second application of the concentration phenomena gives

$$\operatorname{Var}((\operatorname{tr}_m \otimes \operatorname{tr}_n)(f(S_n))) \leq \frac{\pi^2}{8n^2}\mathbb{E}((\operatorname{tr}_m \otimes \operatorname{tr}_n)(f'(S_n)^2)) \tag{4.8}$$

where Var denotes the variance. Now let g be a $C^\infty(\mathbb{R})$-function with values in $[0,1]$ such that g vanishes on $\sigma(S)$ and g is 1 on the complement of $\sigma(s)+]-\varepsilon, \varepsilon[$.

By applying (4.7) and (4.8) to $f = g - 1$, one gets

$$\mathbb{E}((\mathrm{tr}_m \otimes \mathrm{tr}_n)(g(S_n)) = O(\frac{1}{n^2}) \tag{4.9}$$

$$\mathrm{Var}((\mathrm{tr}_m \otimes \mathrm{tr}_n)g(S_n)) = O(\frac{1}{n^4}). \tag{4.10}$$

By a standard application of the Borel-Cantelli lemma (4.9) and (4.10) imply

$$(\mathrm{tr}_m \otimes \mathrm{tr}_n)(g(S_n(\omega))) = O(n^{-4/3})$$

almost surely. Hence the number of eigenvalues for $S_n(\omega)$ outside $\sigma(s)+]-\varepsilon,\varepsilon[$ is $O(n^{-1/3})$[1] almost surely, but being an integer, the number has to vanish eventually as $n \to \infty$ for almost all $\omega \in \Omega$. Hence (4.3) holds.

5. Other applications of random matrices to C^*-algebras

A C^*-algebra A is called exact if for every short exact sequence of C^*-algebras

$$0 \to J \to B \to B/J \to 0$$

the sequence

$$0 \to A \otimes_{\min} J \to A \otimes_{\min} B \to A \otimes_{\min} (B/J) \to 0$$

is exact (cf. [Ki1], [Wa]). The class of exact C^*-algebras is very large: All nuclear C^*-algebras are exact and the reduced group C^*-algebra $C_r^*(\Gamma)$ is exact for any discrete subgroup Γ of a connected locally compact group (cf. [Ki2]). In 1991 the lecturer proved that 2-quasitraces on unital exact C^*-algebras are traces (cf. [Haa1]). Combined with results of Handelman [Han] and Blackadar and Rrdam [BR], this implies that

Every stably finite exact unital C^*-algebra has a tracial state. (5.1)

Every state on the K_0-group, $K_0(A)$ of an exact unital C^*-algebra A is induced by a tracial state on A. (5.2)

Later, Thorbjrnsen and the lecturer found new proofs based on random matrices for (5.1) and (5.2). The key step in the proof was to show:

Theorem 5.1 [HT2] *Let A be an exact unital C^*-algebra, and let $a_1, \ldots, a_r \in A$ be elements in A for which*

$$\sum_{i=1}^{r} a_i^* a_i = c\mathbf{1_A} \qquad \textit{where } c > 1 \tag{5.3}$$

$$\sum_{i=1}^{r} a_i a_i^* \leq \mathbf{1_A} \tag{5.4}$$

[1] tr_m and tr_n are the normalized traces on $M_m(\mathbb{C})$ and $M_n(\mathbb{C})$.

and let $Y_1^{(n)}, \ldots, Y_r^{(n)}$ *be random* $n \times n$*-matrices whose entries are* rn^2 *independent identically distributed complex Gaussian random variables with density* $\frac{n}{\pi}\exp(-n|z|^2)$, $z \in \mathbb{C}$. *Put*

$$S_n = \sum_{i=1}^{r} a_i \otimes Y_i^{(n)} \tag{5.5}$$

and let $\sigma(S_n^* S_n)$ *be the spectrum of* $S_n^* S_n$ *as a function of* $\omega \in \Omega$ *(the underlying probability space). Then for almost all* $\omega \in \Omega$

$$\limsup_{n\to\infty} \max(\sigma(S_n^* S_n)) \leq (\sqrt{c}+1)^2 \tag{5.6}$$

$$\liminf_{n\to\infty} \min(\sigma(S_n^* S_n)) \geq (\sqrt{c}-1)^2 \tag{5.7}$$

The result is a kind of generalization of the results of Geman 1980 [Ge] and Silverstein 1985 [Si] on the asymptotic behaviour of the largest and smallest eigenvalue of a random matrix of Wishart type. The estimates (5.6) and (5.7) were proved by careful moment estimates and lengthy combinatorial arguments. With Theorem 4.1 at hand, a much simpler proof of (5.6) and (5.7) can now be obtained (cf. [HT4]).

Theorem 5.1 is not true in the general non-exact case (cf. [HT3]). It is unknown whether (5.1) or (5.2) hold for general C^*-algebras. Both problems are equivalent to Kaplansky's problem from the 1950's: Is every AW*-factor of type II_1 a von Neumann factor of type II_1?

Let me end this section by discussing another application of Theorem 4.1: Junge and Pisier proved in [JP] that

$$B(H) \otimes_{\max} B(H) \neq B(H) \otimes_{\min} B(H). \tag{5.8}$$

In the proof they consider a sequence of constants $C(k)$, $k \in \mathbb{N}$: For fixed $k \in \mathbb{N}$ $C(k)$ is the infimum of all $C > 0$ for which there exists a sequence of k-tuples of unitary matrices $(u_1^{(m)}, \ldots, u_k^{(m)})_{m\in\mathbb{N}}$ of size $n(m) \in \mathbb{N}$, such that for all $m \neq m'$:

$$\| \sum_{i=1}^{k} u_i^{(m)} \otimes u_i^{(m')} \| \leq C.$$

To obtain (5.8), Junge and Pisier proved that $\lim_{k\to\infty} \frac{C(k)}{k} = 0$. Subsequently, Pisier [P2] proved that $C(k) \geq 2\sqrt{k-1}$ for all $k \in \mathbb{N}$ and Valette [V] proved, using Ramanujan graphs, that $C(k) \leq 2\sqrt{k-1}$ when k is of the form $k = p+1$ for an odd prime number p. It is an easy consequence of Corollary 4.2 that $C(k) \leq 2\sqrt{k-1}$ for all $k \geq 2$ and hence $C(k) = 2\sqrt{k-1}$ for all $k \geq 2$ (see [HT4]).

6. The invariant subspace problem relative to a von Neumann algebra

The invariant subspace problem for operators on general Banach spaces were settled by Enflo [E] and Read [Re] in the 1980's, but for Hilbert spaces the problem is still open:

Problem 6.1 [Hal, pp. 100–101] Let H be a separable infinite dimensional Hilbert space, and let $T \in B(H)$. Does there exist a non-trivial closed T-invariant subspace of H?

More generally, one has the invariant subspace problem relative to a von Neumann algebra:

Problem 6.2 Let $M \subseteq B(H)$ be a von Neumann algebra on a separable Hilbert space H, and let $T \in M$. Does there exist a non-trivial closed T-invariant subspace K for T, such that K is affiliated with M (i.e. K is of the form $K = P(H)$ for a projection $P \in M$)?

The problem is only interesting when $\dim(M) = +\infty$ and when M is a factor, i.e. when the center of M is just $\mathbb{C}1_M$.

The infinite dimensional factors were divided into 4 types by Murray and von Neumann in the late 1930's (cf. [KR, Vol.2]).

Type I_∞: These are isomorphic to $B(K)$ for some infinite dimensional Hilbert space.

Type II_1: M has a tracial state, i.,e. there exists a functional $\mathrm{tr}\colon M \to \mathbb{C}$, such that $\mathrm{tr}(1_M) = 1$, $\mathrm{tr}(S^*S) \geq 0$ and $\mathrm{tr}(ST) = \mathrm{tr}(TS)$ for all $S, T \in M$.

Tupe II_∞: $M \simeq N\widehat{\otimes}B(K)$ where N is type II_1 and $\dim K = +\infty$.

Type III: All other infinite dimensional factors.

In all 4 cases, problem 2 remains open (the Type I_∞ case is of course equivalent to Problem 7.1). We will in the following address the invariant subspace problem relative to a factor of type II_1.

7. The Fuglede-Kadison determinant and Brown's spectral distribution measure

Let M be a II_1-factor. Then M has a unique tracial state tr, and tr is normal and faithful (see eg. [KR, Vol.2, Sect.8]. The *Fuglede-Kadison determinant* $\Delta\colon M \to [0, \infty)$ can be defined (cf. [FK]) by:

$$\Delta(T) = \lim_{\varepsilon \downarrow 0} \exp(\mathrm{tr}(\log(T^*T + \varepsilon 1)^{\frac{1}{2}})), \quad t \in M. \tag{7.1}$$

If T is invertible, one has

$$\Delta(T) = \exp(\mathrm{tr}(\log|T|))$$

where $|T| = (T^*T)^{\frac{1}{2}}$. Moreover Δ has the following properties:

$$\begin{aligned} \Delta(ST) &= \Delta(S)\Delta(T), \quad S, T \in M \\ \Delta(T) &= \Delta(T^*) = \Delta(|T|), \quad T \in M \\ \Delta(U) &= 1, \quad \text{when } U \in M \text{ is unitary.} \end{aligned}$$

Δ is an upper semi-continuous function on M but it is not continuous in the norm-topology on M.

Theorem 7.1 (L.G. Brown 1983 [Br]) *Let M be a II_1-factor and let $T \in M$. Then the function*

$$\varphi\colon \lambda \to \frac{1}{2\pi} \log \Delta(T - \lambda 1), \quad \lambda \in \mathbb{C}$$

is subharmonic and its Laplacian taken in distribution sense

$$\mu_T = \left(\frac{\partial^2}{\partial \lambda_1^2} + \frac{\partial^2}{\partial \lambda_2^2} \right) \varphi \tag{7.2}$$

($\lambda_1 = Re\,\lambda$, $\lambda_2 = Im\,\lambda$) is a probability measure in $\mathbb{C}$ concentrated on the spectrum $\sigma(T)$ of T.

Definition 7.2 *The above measure μ_T is called Brown's spectral distribution measure for T or just the Brown measure for T.*

Example 7.3
a) The Fuglede-Kadison determinant and the Brown measure also make sense for $M = M_n(\mathbb{C})$, and $\mathrm{tr} = \frac{1}{n}\mathrm{Tr}$ the normalized trace on $M_n(\mathbb{C})$. In this case one gets

$$\begin{aligned} \Delta(T) &= \sqrt[n]{|\det T|} \\ \mu_T &= \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}, \end{aligned}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of T repeated according to root multiplicity, and δ_{λ_i} is the Dirac measure at λ_i.
b) If T is a normal operator (i.e. $T^*T = TT^*$) in a factor of type II_1, T has a spectral resolution

$$T = \int_{\sigma(T)} \lambda dE(\lambda).$$

In this case μ_T is equal to $\mathrm{tr} \circ E$.

Methods for computing Brown measures have been developed by Larsen and the lecturer [HL] and by Biane and Lehner [BL].

8. Spectral subspaces for operators in $\mathbf{II_1}$-factors

In 1968, Apostol [Ap] and Foias [Fo1], [Fo2] introduced the notion of spectral subspaces for certain well behaved operators on Banach spaces, the decomposable operators (see [LN] for a modern treatment of this theory):

Definition 8.1 [LN, Definition 1.1.1] *An operator T on a Banach space X is called decomposable if for any open covering $\mathbb{C} = V \cup W$ of the complex plane, there exist closed T-invariant subspaces Y, Z of X such that*

$$X = Y + Z \tag{8.1}$$

$$\sigma(T|_Y) \subseteq V \quad \text{and } \sigma(T|_Z) \subseteq W. \tag{8.2}$$

If $T \in B(X)$ is decomposable, it has a *spectral capacity*, i.e. there exists a map E from the closed subsets of $\mathbb{C}$ into the closed T-invariant subspaces of X, such that

$$\begin{aligned} E(\emptyset) &= 0 \quad \text{and } E(\mathbb{C}) = X && (8.3)\\ X &= E(\bar{V}_1) + \cdots + E(\bar{V}_N) \quad \text{for every finite} && (8.4)\\ &\quad \text{open covering } \mathbb{C} = V_1 \cup V_2 \cup \cdots \cup V_n \\ E(\cap_{n=1}^{\infty} F_n) &= \cap_{n=1}^{\infty} E(F_n), \quad F_n \subseteq \mathbb{C} \text{ closed} && (8.5)\\ \sigma(T_{|E(F)}) &\subseteq F, \quad F \subseteq \mathbb{C} \text{ closed.} && (8.6) \end{aligned}$$

Moreover, a spectral capacity is unique (cf. [LN, Sect.1]).

In this section we will discuss a new method for constructing spectral subspaces of operators which works for all operators in "almost all" II_1-factors, regardless of whether the operator is decomposable in the above sense.

Definition 8.2 *A II_1-factor M on a separable Hilbert space has the* embedding property *if it can be embedded in the ultrapower R^ω of the hyperfinite II_1-factor R for some free ultrafilter ω on the natural numbers.*

All II_1-factors of current interest have this embedding property, and in fact no counterexamples are known. The question whether every II_1-factor on a separable Hilbert space can be embedded in R^ω was first raised by Connes in 1976 [Co] (see also [Ki2] and [HW] for further discussions about this problem).

Let M be a II_1-factor, $M \subseteq B(H)$, and let $T \in M$. If $K \subseteq H$ is a non-trivial closed T-invariant subspace affiliated with M, and $P = P_K$ is the orthogonal projection on M, then according to the decomposition, $H = K \oplus K^\perp$, we can write

$$T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}, \tag{8.7}$$

where $T_{11} = PTP$ and $T_{22} = (1-P)T(1-P)$ are elements of the II_1-factors $M_1 = PMP$ and $M_2 = (1-P)M(1-P)$. Let $\mu_{T_{11}}$ and $\mu_{T_{22}}$ be the Brown measures of T_{11} and T_{22} computed relative to M_1 and M_2 (respectively) then by [Br]:

$$\mu_T = a\mu_{T_{11}} + (1-a)\mu_{T_{22}} \tag{8.8}$$

where $a = \mathrm{tr}_M(P)$.

The main result of [Haa2] is

Theorem 8.3 [Haa2] *Let M be II_1-factor with the embedding property, and let $T \in M$. Then for every Borel set $B \subseteq \mathbb{C}$ there is a unique T-invariant subspace K affiliated with M, such that $\mu_{T_{11}}$ is concentrated on B and $\mu_{T_{22}}$ is concentrated on $\mathbb{C}\backslash B$, where T_{11} and T_{22} are defined as in* (8.7). *Moreover, $\mathrm{Tr}_M(P_K) = \mu_T(B)$, where $P_K \in M$ is the projection onto K.*

Remark 8.4 If T is decomposable and B is closed, then the subspace K coincide with the spectral subspace $E(B)$ characterized by (8.3)–(8.6). However, already in the hyperfinite II_1-factor R, there are operators T which are not decomposable.

Corollary 8.5 [Haa2] *Let $T \in M$, where M is a II_1-factor with the embedding property. If the Brown measure μ_T of T is not concentrated in a single point, then T has a non-trivial closed invariant subspace affiliated with M.*

Remark 8.6 Corollary 8.5 reduced the invariant subspace problem for II_1-factors M with the embedding problem to operators $T \in M$ for which $\mu_T = \delta_0$ (the Dirac-measure at 0). It can be shown that $\mu_T = \delta_0$ if and only if

$$\lim_{n\to\infty} ((T^*)^n T^n)^{\frac{1}{n}} = 0$$

in the strong operator topology on M (cf. [Haa2]).

In the rest of this section, I will briefly outline the proof of Theorem 8.3.

Let M be a II_1-factor and let $T \in M$. Define the modified spectrum $\sigma'(T)$ and modified spectral radius $r'(T)$ by

$$\begin{aligned} \sigma'(T) &= \mathrm{supp}(\mu_T) \\ r'(T) &= \max\{|\lambda| \mid \lambda \in \sigma'(T)\}. \end{aligned}$$

Then $\sigma'(T) \subseteq \sigma(T)$ and $r'(T) \leq r(T)$.

The classical spectral radius formula

$$r(T) = \lim_{n\to\infty} \|T^n\|^{\frac{1}{n}}$$

has a modified version (cf. [Haa2]):

$$r'(T) = \lim_{p\to\infty} (\lim_{n\to\infty} \|T^n\|_{\frac{p}{n}}^{\frac{1}{n}})$$

where $\|S\|_p = \mathrm{tr}_M(|S|^p)^{\frac{1}{p}}$, $p > 0$.

Spectral subspace lemma 8.7 [Haa2] *Let M be a II_1-factor. (Here we do not need the embedding property.) Let $T \in M$ and let $F \subseteq \mathbb{C}$ be a closed set. Then*

(a) There exists a maximal closed T-invariant subspace K affiliated with M such that $\sigma'(T_{|K}) \subseteq F$, where $\sigma'(T_{|K})$ is the modified spectrum of the operator $T_{|K}$ considered as an element of the II_1-factor $P_K M P_K$ (P_K is the projection of H onto K).

(b) Let $K(F)$ be the subspace K defined by (a). Then

$$\mathrm{tr}_M(P_{K(F)}) \leq \mu(F)$$

for all closed subsets F of $\mathbb{C}$.

Random distortion lemma 8.8 [Haa2] *Let M be a II_1-factor with the embedding property and let $T \in M$. Then*

(a) There exist natural numbers $k(1) < k(2) < \ldots$ *and* $T_n \in M_{k(n)}(\mathbb{C})$ *such that*

$$\sup_{n\in\mathbb{N}} \|T_n\| < \infty. \tag{8.9}$$

(b) For every non-commutative polynomial p *in two variables*

$$\lim_{n\to\infty} \mathrm{tr}_{k(n)}(p(T_n, T_n^*)) = \mathrm{tr}(p(T, T^*)) \tag{8.10}$$

where $\mathrm{tr}_{k(n)}$ *is the normalized trace on* $M_n(\mathbb{C})$.

(c) Furthermore, there exists a sequence $T_n' \in M_{k(n)}(\mathbb{C})$ *such that*

$$\lim_{n\to\infty} \|T_n' - T_n\|_p = 0 \quad \textit{for some } p > 0 \tag{8.11}$$

$$\lim_{n\to\infty} \Delta(T_n' - \lambda 1) = \Delta(T - \lambda 1) \quad \textit{for almost all } \lambda \in \mathbb{C} \tag{8.12}$$

$$\lim_{n\to\infty} \mu_{T_n'} = \mu_T \quad \textit{weakly in Prob}(\mathbb{C}). \tag{8.13}$$

The embedding property is needed in (b). To pass from (b) to (c) we use a random distortion argument where we put

$$T_n' = T_n + \varepsilon_n X_n Y_n^{-1}$$

where X_n, Y_n are random Gaussian matrices with independent entries and $\varepsilon_n \to 0$. Subsequently Sniady proved [Sn1] that by using a different random distortion, one can obtain a stronger result, namely in (c), (8.11) can be replaced by

$$\lim_{n\to\infty} \|T_n' - T_n\|_\infty = 0$$

where $\|\cdot\|_\infty$ is the operator norm.

The random distortion lemma is used to reduce the proof of Theorem 8.3 to the case of $M = M_n(\mathbb{C})$ by an ultraproduct argument. For $M = M_n(\mathbb{C})$, Theorem 8.3 is a corollary of Jordan's normal form.

9. Voiculescu's circular operator Y and the strictly upper triangular operator T

Prior to the proof of theorem 8.3, Dykema and the lecturer had constructed invariant subspaces for special operators in factors of type II_1. An example of particular interest is Voiculescu's circular operator Y, which can be written as

$$Y = \frac{1}{\sqrt{2}}(X_1 + iX_2)$$

where (X_1, X_2) is a semicircular system (cf. Section 3.). The von Neumann algebra $M = VN(Y)$ generated by Y is isomorphic to $L(F_2)$ (the von Neumann associated to a free group on two generators) which is a factor of type II_1. The operator Y is far from being normal and for some time it was considered a possible counterexample

for the invariant subspace problem relative to the II_1-factor it generates. In [HL] Larsen and the lecturer proved that

$$\sigma(Y) = \overline{D} \quad \text{(the closed unit disc in } \mathbb{C}) \tag{9.1}$$

$$\text{The Brown measure } \mu_Y \text{ of } Y \text{ is the uniform distribution on } \overline{D}, \text{ i.e. it has constant density } \tfrac{1}{\pi}. \tag{9.2}$$

Theorem 9.1 [DH1] *For each $r \in (0,1)$ there is a unique projection $p \in M = VN(Y)$ such that*

$$pYp = Yp \quad \textit{(i.e. the range of } p \textit{ is } Y\textit{-invariant)} \tag{9.3}$$

$$\sigma(pYp) \subseteq \{z \in \mathbb{C} \mid |z| \leq r\} \tag{9.4}$$

$$\sigma((1-p)Y(1-p)) \subseteq \{z \in \mathbb{C} \mid r \leq |z| \leq 1\} \tag{9.5}$$

where the spectra in (9.4) *and* (9.5) *are computed relative to pMp and $(1-p)M(1-p)$. Moreover*

$$\mathrm{tr}_M(p) = r^2. \tag{9.6}$$

This result was generalized to arbitrary R-diagonal elements by Sniady and Speicher [SS]. Later Dykema and the lecturer proved

Theorem 9.2 [DH2] *Voiculescu's circular operator is decomposable in the sense of Apostol and Foias (see Definition 8.1).*

In [DH2] we also considered the "strictly upper triangular operator" T. It is defined in terms of its random matrix model:

Theorem/Definition 9.3 [DH2] *Let for each $n \in \mathbb{N}$ T_n denote the strictly upper triangular random matrix*

$$T_n = \begin{pmatrix} 0 & t_{11}^{(n)} & \cdots & t_{1n}^{(n)} \\ \ddots & & \ddots & t_{n-1,n}^{(n)} \\ 0 & & & 0 \end{pmatrix} \tag{9.7}$$

*for which the entries $(t_{ij}^{(n)})_{i<j}$ are $\frac{n(n-1)}{2}$ independent identically distributed complex Gaussian random variables with densities $\frac{n}{\pi}\exp(-n|z|^2)$, $z \in \mathbb{C}$. Then there is an operator T in a II_1-factor M such that T_n converges in *-moments to T, i.e.*

$$\mathrm{tr}_M(P(T,T^*)) = \lim_{n\to\infty} \mathbb{E}\mathrm{tr}_n(P(T_n,T_n^*)) \tag{9.8}$$

for every non-commutative polynomial P. T is called the strictly upper triangular operator.

The strictly upper triangular operator is quasi nilpotent, i.e. $\sigma(T) = \{0\}$, and therefore its Brown measure μ_T is equal to δ_0. In view of remark 8.6 it could be a candidate for a counterexample to the invariant subspace problem relative to a II_1-factor. However, this is not the case:

Dykema and the lecturer proved in [DH2] that

$$\mathrm{tr}((T^*T)^n) = \frac{n^n}{(n+1)!}, \qquad n \in \mathbb{N} \tag{9.9}$$

and in [Sn2], Sniady proved

$$\mathrm{tr}(((T^k)^*T^k)^n) = \frac{n^{nk}}{(nk+1)!}, \qquad n,k \in \mathbb{N}, \tag{9.10}$$

a formula which was conjectured in [DH2].

Based on (9.10) and its proof, we recently proved

Theorem 9.4 [DH3] *Let T be as above. Put $S_k = k((T^k)^kT^k)^{\frac{1}{k}}$ and let $F\colon [0,\pi] \to [0,1]$ be the strictly increasing function given by $F(0) = 0$, $F(\pi) = 1$ and*

$$F\left(\frac{\sin v}{v}\exp(v\cot v)\right) = 1 - \frac{v}{\pi} + \frac{1}{\pi}\frac{\sin^2 v}{v}, \quad 0 < v < \pi. \tag{9.11}$$

Then $F(S_k)$ converges in strong operator topology to the "diagonal operator" D_0 with matrix model

$$D_{0,n} = \begin{pmatrix} \frac{1}{n} & & & 0 \\ & \frac{2}{n} & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}. \tag{9.12}$$

In particular $D_0 \in VN(T)$. Moreover $VN(T)$ is isomorphic to $L(F_2)$ and the ranges of the projections $1_{[0,t]}(D_0)$, $0 < t < 1$, form an uncountable family of non-trivial invariant subspaces for T affiliated with $VN(T)$.

References

[An] J. Anderson, A C^*-algebra A for which $\mathrm{Ext}(A)$ is not a group, *Annals Math.* 107 (1978), 455–458.

[Ap] C. Apostol, Spectral decomposition and functional calculus, *Rev. Roum. Math. Pures Appl.* 13 (1968), 1481–1528.

[Ar] L. Arnold, On the asymptotic distribution of the eigenvalues of random matrices, *Journ. Math. Anal. Appl.* 20 (1967), 262–268.

[Ba] Z.D. Bai, Methodologies in spectral analysis of large dimensional random matrices, A review, *Statistica Sinica* 9 (1999), 611–677.

[Bl] B. Blackadar, K-theory for operator algebras, *Math. Sci. Res. Inst. Publ.* 5, Springer Verlag (1986).

[BDF] L.G. Brown, R.G. Douglas and P.A. Fillmore, Extensions of C^*-algebras and K-homology, *Ann. of Math.* 105 (1977), 265–324.

[BK] B. Blackadar, E. Kirchberg, Generalized inductive limits of finite dimensional C^*-algebras, *Math. Ann.* 307 (1997), 343–380.

[BL] P. Biane and F. Lehner, Computation of some examples of Brown's spectral measure in free probability theory. *Colloqium Mathematicum* 90 (2001), 181–211.

[BR] B. Blackadar and M. Rrdam, Extending states on preordered semigroups and the existence of quasitraces on C^*-algebras, *Journ. Algebra* 152 (1992), 240–247.

[Br] L.G. Brown, Lidskii's theorem in the type II case, Geometric methods in operator algebras (Kyoto 1983), H. Araki and E. Effros (Eds.) *Pitman Res. notes in Math.* Ser 123, Longman Sci. Tech.(1986), 1–35.

[BY] Z.D. Bai and Y.Q. Yin, Neccesary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *Anal. of Probab.* 16, 1729–1741 (1988).

[CE] M.D. Choi and E. Effros, The completely positive lifting problem for C^*-algebras, *Ann. of Math.* 104 (1976), 585–609.

[Co] A. Connes, Classification of injective factors, *Ann. of Math.* 104 (1976), 73–115.

[DH1] K. Dykema and U. Haagerup, Invariant subspaces of Voiculescu's circular operator, *Geom. Funct. Anal.* 11 (2001), 693–741.

[DH2] K. Dykema and U. Haagerup, DT-operators and decomposability of Voiculescu's circular operator. Preprint 2002, http://xxx.arxiv.org/abs/math.OA/0205077.

[DH3] K. Dykema and U. Haagerup, In preparation.

[E] P. Enflo, On the invariant subspace problem for Banach Spaces, *Acta Math.* 158 (1987), 213–313.

[EH] E. Effros and U. Haagerup, Lifting problems and local reflexivity for C^*-algebras, Duke Math. Journ. 52 (1985), 103–128.

[Fo1] C. Foias, Spectral maximal spaces and decomposable operators on Banach spaces, *Arch. Math.* 14 (1963), 341–349.

[Fo2] C. Foias, Spectral capacities and decomposable operators, *Rev. Roum. Math. Pures Appl.* 13 (1968), 1539–1545.

[FK] B. Fuglede and R.V. Kadison, Determinant theory in finite factors, *Ann. Math.* 55 (1952), 520–530.

[Ge] S. Geman, A limit theorem for the norm of random matrices, *Annals Prob.* 8 (1980) 252–261.

[Gi] J. Ginibre, Statistical ensembles of complex, quaternionic and real matrices, *Journ. Math. Phys.* 6 (1965), 440–449.

[Haa1] U. Haagerup, Qusitraces on exact C*-algebras are traces, Unpublished manuscript (1991).

[Haa2] U. Haagerup, Spectral decomposition of all operators in a II_1-factor, which is embeddable in R^ω (Preliminary version). MSRI 2001.

[Hal] P.R. Halmos, A Hilbert space problem book, *2nd Ed. Graduate Texts in Mathematics* 19, Springer Verlag 1982.

[Han] D. Handelman, Homomorphisms of C^*-algebras to finite AW^*-algebras, *Michigan Math. Journ.* 28 (1991), 229–240.

[HL] U. Haagerup and F. Larsen, Brown's spectral distribution measure for R-

diagonal elements in finite von Neumann algebras, *Journ. Funct. Analysis* 176 (2000), 331–367

[HP1] F. Hiai and D. Petz, The semicircle law, Free random variables and entropy, *Amer. Math. Soc.* 2000.

[HP2] F. Hiai and D. Petz, Asymptotic freeness almost everywhere for random matrices, *Acta Sci. Math.* (Szeged) 66 (2000), 809–834.

[HT1] U. Haagerup and S. Thorbjrnsen, Random matrices with complex Gaussian entries. Preprint 1998.

[HT2] U. Haagerup and S. Thorbjrnsen, Random matrices and K-theory for exact C^*-algebras. *Documenta Math.* 4 (1999), 330–441.

[HT3] U. Haagerup and S. Thorbjrnsen, Random Matrices and non-exact C^*-algebras, "C^*-algebras" (J. Cuntz, S. Echterhoff ed.) (2000), 71–91.

[HT4] U. Haagerup and S. Thorbjrnsen, A new application of random matrices: $\mathrm{Ext}(C_r^*(F_2))$ is not a group. Preprint 2002.

[HW] U. Haagerup and C. Winslw, The Effros-Marshal topology in the space of von Neumann algebras II, *Journ. Funct. Anal.* 171 (2000), 401–431.

[JP] M. Junge and G. Pisier, Bilinear forms on exact operator spaces and $B(H) \otimes B(H)$, *Geom. Funct. Analysis* 5 (1995), 329–363.

[Ki1] E. Kirchberg, The Fubini Theorem for Exact C^*-algebras, *Journ. Operator Theory* 10 (1983), 3–8.

[Ki2] E. Kirchberg, On non-semisplit extensions, tensor products and exactness of group C^*-algebras, *Invent. Math.* 112 (1993), 449–489.

[KR] R.V. Kadison and J. Ringrose, Fundamentals of the theory of operator algebras, Vol. II, *Academic Press* 1986.

[LN] K.B. Laursen and M.M. Neumann, An introduction to Local spectral theory, Clarendon Press, Oxford 2000.

[Le] F. Lehner, Computing norms of free operators with matrix coefficients. *Amer. J. Math.* 121 (1999), 453–486.

[Me] M.L. Mehta, Random matrices, second edition, *Academic Press* (1991).

[Mi] V.D. Milman, The concentration phenomenon of finite dimensional normed spaces, *Proc. International Congr. Math.*, Berkeley, vol 2 (1987), 961–975.

[P1] G. Pisier, The volume of convex bodies and Banach space geometry, *Cambridge University Press* (1989).

[P2] G. Pisier, Quadratic forms in unitary operators, *Linear algebra and its Appl.* 267 (1997),125–137.

[Re] C.J. Reed, A solution to the invariant subspace problem, *Bull. London Math. Soc.* 16 (1984), 337–401.

[Ro] J. Rosenberg, Quasidiagonality and Nuclearity (appendix to strongly quasidiagonal operators by D. Hadwin), *Journ. Operator Theory* 18 (1987), 15–18.

[Si] J.W. Silverstein, The smallest eigenvalue of a large dimensional Wishart matrix, *Annals Prob.* 13 (1985), 1364–1368.

[Sn1] P. Sniady, Random regularization of Browns spectral measure, *Journ. Funct. Analysis* 193 (2002), 291–313.

[Sn2] P. Sniady, Multinomial identities arising from the free probability, preprint

2002.

[SS] P. Sniady and R. Speicher, Continuous family of invariant subspaces for R-diagonal operators, *Invent. Math.* 146 (2001), 329–363.

[T] S. Thorbjrnsen, Mixed moments of Voiculescu's Gaussian random matrices, *Journ. Funct. Anal.* 176 (2000), 213–246.

[TW1] C. Tracy and H. Widom, Level-spacing distributions and the Airy kernel, *Comm. Math. Phys.* 159 (1994), 151–174.

[TW2] C. Tracy and H. Widom, On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* 177 (1996), 727–754.

[V] A. Valette, An application of Ramanujan graphs to C^*-algebra tensor products, *Discrete Math.* 167 (1997), 597–603.

[V1] D. Voiculescu, A non commutative Weyl-von Neumann Theorem, *Rev. Roum. Pures et Appl.* 21 (1976), 97–113.

[V2] D. Voiculescu, Symmetries of some reduced free group C^*-algebras, "Operator Algebras and Their Connections with Topology and Ergodic Theory", *Lecture Notes in Math.* Vol. 1132, Springer-Verlag 1985, 556–588.

[V3] D. Voiculescu, Circular and semicircular systems and free product factors, "Operator Algebras, Unitary representations, Algebras, and Invariant Theory", *Progress in Math.* Vol. 92, Birkhuser, 1990, 45–60.

[V4] D. Voiculescu, Limit laws for random matrices and free products, *Inventiones Math.* 104 (1991), 202–220.

[V5] D. Voiculescu, A note on quasi-diagonal C^*-algebras and homotopy. *Duke Math. Journ.* 62 (1991), 267–271.

[V6] D. Voiculescu, Around quasidiagonal operators. *Integral Equations and Operator Theory* 17 (1993), 137–149.

[V7] D. Voiculescu, Operators on certain non-commutative random variables, "Recent advances in operator algebras, Orleans 1992" *Asterisque* 232 (1995), 243–275.

[V8] D. Voiculescu, Free Probability Theory: Random Matrices and von Neumann algebras, *Proceedings of the International Congress of Mathematicians*, Zürich 1994, Birkhäuser Verlag, Basel 1995.

[VDN] D. Voiculescu, K. Dykema and A. Nica, Free Random Variables, CMR Monograph Series 1, *American Mathematical Society*, 1992.

[Wa] S. Wassermann, Exact C*-algebras and related topics, *Seoul National University Lecture Notes Series* 19 (1994).

[Wi] E. Wigner, Characterictic vectors of boardered matrices with infinite dimensions, *Ann. Math.* 62 (1955), 548–564.

ICM 2002 · Vol. I · 291–317

Algebraic Topology and Modular Forms

M. J. Hopkins*

1. Introduction

The problem of describing the homotopy groups of spheres has been fundamental to algebraic topology for around 80 years. There were periods when specific computations were important and periods when the emphasis favored theory. Many mathematical invariants have expressions in terms of homotopy groups, and at different times the subject has found itself located in geometric topology, algebra, algebraic K-theory, and algebraic geometry, among other areas.

There are basically two approaches to the homotopy groups of spheres. The oldest makes direct use of geometry, and involves studying a map $f : S^{n+k} \to S^n$ in terms of the inverse image $f^{-1}(x)$ of a regular value. The oldest invariant, the degree of a map, is defined in this way, as was the original definition of the Hopf invariant. In the 1930's Pontryagin[1] [43, 42] showed that the homotopy class of a map f is completely determined by the geometry of the inverse image $f^{-1}(B_\epsilon(x))$ of a small neighborhood of a regular value. He introduced the basics of framed cobordism and framed surgery, and identified the group $\pi_{n+k}S^n$ with the cobordism group of smooth k-manifolds embedded in $\mathbb{R}^{n+k}$ and equipped with a framing of their stable normal bundles.

The other approach to the homotopy groups of spheres involves comparing spheres to spaces whose homotopy groups are known. This method was introduced by Serre [50, 51, 16, 15] who used Eilenberg-MacLane spaces $K(A, n)$, characterized by the property

$$\pi_i K(A,n) = \begin{cases} A & i = n \\ 0 & \text{otherwise.} \end{cases}$$

By resolving a sphere into Eilenberg-MacLane spaces Serre was able to compute $\pi_{k+n}S^n$ for all $k \leq 8$.

For some questions the homotopy theoretic methods have proved more powerful, and for others the geometric methods have. The resolutions that lend themselves to computation tend to use spaces having convenient homotopy theoretic properties, but with no particularly accessible geometric content. On the other hand, the geometric methods have produced important homotopy theoretic moduli spaces and

*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: mjh@math.mit.edu

[1]Lefschetz reported on this work at the 1936 ICM in Oslo.

relationships between them that are difficult, if not impossible, to see from the point of view of homotopy theory. This metaphor is fundamental to topology, and there is a lot of power in spaces, like the classifying spaces for cobordism, that directly relate to both geometry and homotopy theory. It has consistently proved important to understand the computational aspects of the geometric devices, and the geometric aspects of the computational tools.

A few years ago Haynes Miller and I constructed a series of new cohomology theories, designed to isolate certain "sectors" of computation. These were successful in resolving several open issues in homotopy theory and in contextualizing many others. There seemed to be something deeper going on with one of them, and in [27] a program was outlined for constructing it as a "homotopy theoretic" moduli space of elliptic curves, and relating it to the Witten genus. This program is now complete, and we call the resulting cohomology theory tmf (for *topological modular forms*). The theory of topological modular forms has had applications in homotopy theory, in the theory of manifolds, in the theory of lattices and their θ-series, and most recently seems to have an interesting connection with the theory of p-adic modular forms. In this note I will explain the origins and construction of tmf and the way some of these different applications arise.

2. Sixteen homotopy groups

By the Freudenthal suspension theorem, the value of the homotopy group $\pi_{n+k}S^n$ is independent of n for $n > k+1$. This group is k^{th} stable homotopy group of the sphere, often written $\pi_k^{\text{st}}(S^0)$, or even as $\pi_k S^0$ if no confusion is likely to result. In the table below I have listed the values of $\pi_{n+k}S^n$ for $n \gg 0$ and $k \leq 15$.

k	0	1	2	3	4	5	6	7	8
$\pi_{n+k}S^n$	$\mathbb{Z}$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/24$	0	0	$\mathbb{Z}/2$	$\mathbb{Z}/240$	$\mathbb{Z}/2 \oplus \mathbb{Z}/2$

9	10	11	12	13	14	15
$(\mathbb{Z}/2)^3$	$\mathbb{Z}/6$	$\mathbb{Z}/504$	0	$\mathbb{Z}/3$	$(\mathbb{Z}/2)^2$	$\mathbb{Z}/2 \oplus \mathbb{Z}/480$

In geometric terms, the group $\pi_{n+k}S^n$ is the cobordism group of stably framed manifolds, and a homomorphism from $\pi_{n+k}S^n$ to an abelian group A is a cobordism invariant with values in A. The groups in the above table thus represent universal invariants of framed cobordism. Some, but not all of these invariants have geometric interpretations.

When $k = 0$, the invariant is simply the number of points of the framed 0-manifold. This is the geometric description of the degree of a map.

When $k = 1$ one makes use of the fact that any closed 1-manifold is a disjoint union of circles. The $\mathbb{Z}/2$ invariant is derived from the fact that a framing on S^1 differs from the framing which bounds a framing of D^2 by an element of $\pi_1 SO(N) = \mathbb{Z}/2$.

There is an interesting history to the invariant in dimension 2. Pontryagin originally announced that the group $\pi_{n+2}S^n$ is trivial. His argument made use of the classification of Riemann surfaces, and a new geometric technique, now known as framed surgery. He later [44] correctly evaluated this group, but for his corrected argument didn't need the technique of surgery. Surgery didn't reappear in again until around 1960, when it went on to play a fundamental role in geometric topology. The invariant is based on the fact that a stable framing of a Riemann surface Σ determines a quadratic function $\phi : H^1(\Sigma; \mathbb{Z}/2) \to \mathbb{Z}/2$ whose underlying bilinear form is the cup product. To describe ϕ, note that each 1-dimensional cohomology class $x \in H^1(\Sigma)$ is Poincaré dual to an oriented, embedded 1-manifold, C_x, which inherits a framing of its stable normal bundle from that of Σ. The manifold C_x defines an element of $\pi_1^{\mathrm{st}} S^0 = \mathbb{Z}/2$, and the value of $\phi(x)$ is taken to be this element. The cobordism invariant in dimension 2 is the Arf invariant of ϕ.

A similar construction defines a map

$$\pi_{4k+2}^{\mathrm{st}} S^0 \to \mathbb{Z}/2. \tag{2.1}$$

In [14] Browder interpreted this invariant in homotopy theoretic terms, and showed that it can be non-zero only for $4k + 2 = 2^m - 2$. It is known to be non-zero for $\pi_2^{\mathrm{st}} S^0$, $\pi_6^{\mathrm{st}} S^0$, $\pi_{14}^{\mathrm{st}} S^0$, $\pi_{30}^{\mathrm{st}} S^0$ and $\pi_{62}^{\mathrm{st}} S^0$. The situation for $\pi_{2^m-2}^{\mathrm{st}} S^0$ with $m > 6$ is unresolved, and remains an important problem in algebraic topology. More recently, the case $k = 1$ of (2.1) has appeared in M-theory [58]. Building on this, Singer and I [26] offer a slightly more analytic construction of (2.1), and relate it to Riemann's θ-function.

Using K-theory, Adams [2] defined surjective homomorphisms (the d and e-invariants)

$$\begin{aligned} \pi_{4n-1} S^0 &\to \mathbb{Z}/d_n, \\ \pi_{8k} S^0 &\to \mathbb{Z}/2, \\ \pi_{8k+1} S^0 &\to \mathbb{Z}/2 \oplus \mathbb{Z}/2, \\ \pi_{8k+2} S^0 &\to \mathbb{Z}/2. \end{aligned}$$

where d_n denotes the denominator of $B_{2n}/(4n)$. He (and Mahowald [35]) showed that they split the inclusion of the image of the J-homomorphism, making the latter groups summands. A geometric interpretation of these invariants appears in Stong [53] using Spin-cobordism, and an analytic expression for the e-invariant in terms of the Dirac operator appears in the work of Atiyah, Patodi and Singer [7, 8]. The d-invariants in dimensions $(8k + 1)$ and $(8k + 2)$ are given by the mod 2 index of the Dirac operator [11, 10].

This more or less accounts for the all of the invariants of framed cobordism that can be constructed using known geometric techniques. In every case the geometric invariants represent important pieces of mathematics. What remains is the following list of homotopy theoretic invariants having no known geometric interpretation:

$\cdots$	8	9	10	11	12	13	14	15
$\cdots$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/3$			$\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$

This part of homotopy theory is not particularly exotic. In fact it is easy to give examples of framed manifolds on which the geometric invariants vanish, while the homotopy theoretic invariants do not. The Lie groups $SU(3)$, $U(3)$, $Sp(2)$, $Sp(1) \times Sp(2)$, G_2, $U(1) \times G_2$ have dimensions 8, 9, 10, 13, 14, and 15, respectively. They can be made into framed manifolds using the left invariant framing, and in each case the corresponding invariant is non-zero. We will see that the theory of topological modular forms accounts for *all* of these invariants, and in doing so relates them to the theory of elliptic curves and modular forms. Moreover many new invariants are defined.

3. Spectra and stable homotopy

In order to explain the theory of topological modular forms it is necessary to describe the basics of stable homotopy theory.

3.1. Spectra and generalized homology

Suppose that X is an $(n-1)$-connected pointed space. By the Freudenthal Suspension Theorem, the suspension homomorphism

$$\pi_{n+k}(X) \to \pi_{n+k+1}\Sigma X$$

is an isomorphism in the range $k < 2n-1$. This is the *stable range* of dimensions, and in order to isolate it and study only and *stable homotopy theory* one works in the category of *spectra.*

Definition 3.2 (see [34, 23, 19, 20, 4]) *A* spectrum E *consists of a sequence of pointed spaces* E_n, $n = 0, 1, 2, \ldots$ *together with maps*

$$s_n^E : \Sigma E_n \to E_{n+1} \tag{3.3}$$

whose adjoints

$$t_n^E : E_n \to \Omega E_{n+1} \tag{3.4}$$

are homeomorphisms.

A map $E \to F$ of spectra consists of a collection of maps

$$f_n : E_n \to F_n$$

which is compatible with the structure maps t_n^E and t_n^F.

For a spectrum $E = \{E_n, t_n\}$ the value of the group $\pi_{n+k}E_n$ is independent of n, and is written $\pi_n E$. Note that this makes sense for any $n \in \mathbb{Z}$. More generally, for any pointed space X, the E-homology and E-cohomology groups of X are defined as

$$E^k(X) = [\Sigma^n X, E_{n+k}],$$
$$E_k(X) = \varinjlim \pi_{n+k}E_n \wedge X.$$

Any homology theory is represented by a spectrum in this way, and any map of homology theories is represented (not necessarily uniquely) by a map of spectra. For example, the spectrum HA with HA_n the Eilenberg-MacLane space $K(A,n)$ represents ordinary homology with coefficients in an abelian group A.

3.2. Suspension spectra and Thom spectra

In practice, spectra come about from a sequence of spaces X_n and maps $t_n : \Sigma X_n \to X_{n+1}$. If each of the maps t_n is a closed inclusion, then the collection of spaces

$$(LX)_k = \varinjlim \Omega^n X_{n+k}$$

forms a spectrum. In case $X_n = S^n$, the resulting spectrum is the *sphere spectrum* and denoted S^0. By construction

$$\pi_k S^0 = \pi_k^{\mathrm{st}} S^0 = \pi_{n+k} S^n \qquad n \gg 0.$$

In case $X_n = \Sigma^n X$, the resulting spectrum is the *suspension spectrum* of X, denoted $\Sigma^\infty X$ (or just X when no confusion with the space X is likely to occur). Its homotopy groups are given by

$$\pi_k \Sigma^\infty X = \pi_k^{\mathrm{st}} X = \pi_{n+k} \Sigma^n X \qquad n \gg 0,$$

and referred to as the *stable homotopy groups of* X.

Another important class of spectra are *Thom spectra.* Let $BO(n)$ denote the Grassmannian of n-planes in $\mathbb{R}^\infty$, and $MO(n)$ the Thom complex of the universal n-plane bundle over $BO(n)$. The natural maps

$$\Sigma MO(n) \to MO(n+1)$$

lead to a spectrum MO, the unoriented bordism spectrum. This spectrum was introduced by Thom [54], who identified the group $\pi_k MO$ with the group of cobordism classes of k-dimensional unoriented smooth manifolds. Using the complex Grassmannian instead of the real Grassmannian leads to the *complex cobordism* spectrum MU. The group $\pi_k MU$ can be interpreted as the group of cobordism classes of k-dimensional stably almost complex manifolds [39]. More generally, a Thom spectrum X^ζ is associated to any map

$$\zeta : X \to BG$$

from a space X to the classifying space BG for stable spherical fibrations.

The groups $\pi_k MO$ and $\pi_k MU$ have been computed [54, 39], as many other kinds of cobordism groups. The spectra representing cobordism are among the few examples that lend themselves to both homotopy theoretic and geometric investigation.

3.3. Algebraic structures and spectra

The set of homotopy classes of maps between spectra is an abelian group, and in fact the category of abelian groups makes a fairly good guideline for contemplating the general structure of the category of spectra. In this analogy, spaces correspond to sets, and spectra to abelian groups. The smash product of pointed spaces

$$X \wedge Y = X \times Y/(x, *) \sim (*, y)$$

leads to an operation $E \wedge F$ on spectra analogous to the tensor product of abelian groups. Using this "tensor structure" one can imitate many constructions of algebra in stable homotopy theory, and form analogues of associative algebras (A_∞-ring spectra), commutative algebras (E_∞-ring spectra), modules, etc. The details are rather subtle, and the reader is referred to [20] and [29] for further discussion.

The importance of refining common algebraic structures to stable homotopy theory has been realized by many authors [38, 20, 55, 56], and was especially advocated by Waldhausen.

The theory of topological modular forms further articulates this analogy. It is built on the work of Quillen relating formal groups and complex cobordism. In [45], Quillen portrayed the complex cobordism spectrum MU as the universal cohomology theory possessing Chern classes for complex vector bundles (a *complex oriented* cohomology theory). These generalized Chern classes satisfy a Cartan formula expressing the Chern classes of a Whitney sum in terms of the Chern classes of the summands. But the formula for the Chern classes of a tensor product of line bundles is more complicated than usual one. Quillen showed [45, 4] that it is as complicated as it can be. If E is a complex oriented cohomology theory, then there is a unique power series

$$F[s, t] \in \pi_* E[\![s, t]\!]$$

with the property that for two complex line bundles L_1 and L_2 one has

$$c_1(L_1 \otimes L_2) = F[c_1(L_1), c_1(L_2)].$$

The power series $F[s, t]$ is a formal group law over $\pi_* E$. Quillen showed that when $E = MU$, the resulting formal group law is universal in the sense that if F is any formal group law over a ring R, then there is a unique ring homomorphism $MU_* \to R$ classifying F. In this way the complex cobordism spectrum becomes a topological model for the moduli space of formal group laws.

3.4. The Adams spectral sequence

There are exceptions, but for the most part what computations can be made of the stable and unstable homotopy groups involve approximating a space by the spaces of a spectrum E whose homotopy groups are known, or at least qualitatively understood. A mechanism for doing this was discovered by Adams [1] in the case $E = H\mathbb{Z}/p$, and later for a general cohomology theory by both Adams [3, 5] and Novikov [41, 40]. The device is known as the *E-Adams spectral sequence for X*,

or, in the case $E = MU$, the *Adams-Novikov spectral sequence for* X (or, in case $X = S^0$, just the Adams-Novikov spectral sequence).

The Adams-Novikov spectral sequence has led to many deep insights in algebraic topology (see, for example, [47, 48] and the references therein). It is usually displayed in the first quadrant, with the groups contributing to $\pi_k S^0$ all having x-coordinate k. The y-coordinate is the MU-Adams filtration, and can be described as follows: a stable map $f : S^k \to S^0$ has filtration $\geq s$ if there exists a factorization

$$S^k = X_0 \to X_1 \to \cdots \to X_{s-1} \to X_s = S^0$$

with the property that each of the maps $MU_* X_n \to MU_* X_{n+1}$ is zero. There is a geometric interpretation of this filtration: a framed manifold M has filtration $\geq s$ if it occurs as a codimension n corner in a manifold N with corners, equipped with suitable almost complex structures on its faces (see [32]). The Adams-Novikov spectral starts with the purely algebraic object

$$E_2^{s,t} = \mathrm{Ext}^{s,t}_{MU_* MU}(MU_*, MU_*)\,.$$

The quotient of the subgroup of $\pi^k S^0$ consisting of elements of Adams-Novikov filtration at least s, by the subgroup of those of filtration at least $(s+1)$ is a sub-quotient of the group $\mathrm{Ext}^{s,t}$ with $(t-s) = k$.

3.5. Asymptotics

For a number k, let $g(k) = s$ be the largest integer s for which $\pi_k S^0$ has a non-zero element of Adams-Novikov filtration s. The graph of g is the *MU-vanishing curve*, and the main result of [18] is equivalent to the formula

$$\lim_{k\to\infty} \frac{g(k)}{k} = 0.$$

This formula encodes quite a bit of the large scale structure of the category of spectra (see[18, 28, 48]), and it would be very interesting to have a more accurate asymptotic expression. This is special to complex cobordism. In the case of the original Adams spectral sequence for a finite CW complex X (based on ordinary homology with coefficients in $\mathbb{Z}/p$), it can be shown [25] that

$$\lim_{k\to\infty} \frac{g_X^{H\mathbb{Z}/p}(k)}{k} = \frac{1}{2(p^m - 1)},$$

for some m. This integer m is an invariant of X know as the "type" of X. It coincides with the largest value m for which the Morava K-group $K(m)_* X$ is non-zero. For more on the role of this invariant in algebraic topology, see [28, 22].

Now the E_2-term of the Adams-Novikov spectral sequence is far from being zero above the curve $g(n)$, and a good deal of what happens in spectral sequence has to do with getting rid of what is up there. A few years ago, Haynes Miller, and I constructed a series of spectra designed to classify and capture the way this happens. We were motivated by connective KO-theory, whose Adams-Novikov

spectral sequence more or less coincides with the Adams-Novikov spectral for the sphere above a line of slope $1/2$, and is very easy to understand below that line (and in fact connective KO can be used to capture everything above a line of slope $1/5$ [37, 33, 36]). By analogy we called these cohomology theories EO_n. These spectra were used to solve several problems about the homotopy groups of spheres.

The theory we now call tmf was originally constructed to isolate the "slope $1/6^{\text{th}}$-sector" of the Adams Novikov spectral sequence, and in [?], for the reasons mentioned above, it was called eo_2. In the next section the spectrum tmf will be constructed as a topological model for the moduli space (stack) of generalized elliptic curves.

4. tmf

4.1. The algebraic theory of modular forms

Let C be the projective plane curve given by the Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{4.5}$$

over the ring

$$A = \mathbb{Z}[a_1, a_2, a_3, a_4, a_6].$$

Let

$$A_* = \bigoplus_{n\in\mathbb{Z}} A_{2n},$$

be the graded ring with

$$A_{2n} = H^0(C; (\Omega^1_{C/A})^{\otimes n}).$$

If $u \in A_2$ is the differential

$$u = \frac{dx}{2y + a_1x + a_3},$$

then

$$A_* \approx A[u^{\pm 1}].$$

The A-module A_2 is free over A of rank 1, and is the module of sections of the line bundle

$$\omega := H^0\left(\mathcal{O}_C(-e)/\mathcal{O}_C(-2e)\right) \approx p_*\Omega^1 C.$$

In this expression $p : C \to \operatorname{Spec} A$ is the structure map, and $e : \operatorname{Spec} A \to C$ is the point at ∞.

Let G be the algebraic group of projective transformations

$$x \mapsto \lambda^2 x + r,$$
$$y \mapsto \lambda^3 y + sx + t.$$

Such a transformation carries C to the curve C' defined by an equation

$$y^2 + a_1'xy + a_3'y = x^3 + a_2'x^2 + a_4'x + a_6',$$

for some a_i'. This defines an action of G on A_*. The ring of invariants

$$H^0(G; A_*)$$

is the ring of *modular forms over* $\mathbb{Z}$.

The structure of $H^0(G; A_*)$ was worked out by Tate (see Deligne [17]). After inverting 6 and completing the square and the cube, equation (4.5) can be put in the form

$$\tilde{y}^2 = \tilde{x}^3 + \tilde{c}_4\,\tilde{x} + \tilde{c}_6 \qquad \tilde{c}_4, \tilde{c}_6 \in A[\tfrac{1}{6}],$$

with

$$\tilde{x} = x + \frac{a_1^2 + 4a_2}{12} \qquad \tilde{y} = y + \frac{a_1 x + a_3}{2}.$$

The elements

$$c_4 = 48\,u^4\tilde{c}_4,$$
$$c_6 = 864\,u^6\tilde{c}_6,$$

lie in A_*, and

$$H^0(G; A_*) = \mathbb{Z}[c_4, c_6, \Delta]/(c_4^3 - c_6^2 = 1728\Delta).$$

We'll write

$$M_n = H^0(G; A_{2n})$$

for the homogeneous part of degree $2n$. It is the group of *modular forms of weight* n *over* $\mathbb{Z}$.

4.2. The topological theory of modular forms

In [27, 24] it is shown that this algebraic theory refines from rings to *ring spectra*, leading to a topological model for the theory of elliptic curves and modular forms. Here is a rough idea of how it goes.

The set of regular points of C has a unique group structure in which the point at ∞ is the identity element, and in which collinear points sum to zero. Expanding the group law in terms of the coordinate $t = x/y$ gives a formal group law

$$C^f[s, t] \in A[\![s, t]\!]$$

over A, which, by Quillen's theorem (see §3.3.) is classified by a graded ring homomorphism

$$MU_* \to A_*.$$

The functor

$$X \mapsto MU_*(X) \otimes_{MU_*} A_*$$

is not quite a cohomology theory, but it becomes one after inverting c_4 or Δ.

Based on this, a spectrum E_A can be constructed with

$$\pi_* E_A = A_*,$$

and representing a complex oriented cohomology theory in which the formula for the first Chern class of a tensor product of complex line bundles is given by

$$c_1(L_1 \otimes L_2) = u^{-1} C^f (u\, c_1(L_1), u\, c_1(L_2)) .$$

A spectrum E_G can be constructed out of the affine coordinate ring of G in a similar fashion, as can an "action" of E_G on E_A. The spectrum tmf is defined to be the (-1)-connected cover of the homotopy fixed point spectrum of this group action.

To actually carry this out requires quite a bit of work. The difficulty is that the theory just described only defines an action of E_G on E_A up to homotopy, and this isn't rigid enough to form the homotopy fixed point spectrum. In the end it can be done, and there turns out to be an unique way to do it.

4.3. The ring of topological modular forms

The spectrum tmf is a homotopy theoretic refinement of the ring $H^0(G; A_*)$, there is a spectral sequence

$$H^s (G; A_t) \Rightarrow \pi_{t-s} \operatorname{tmf} .$$

The ring $\pi_* \operatorname{tmf}$ is the ring of *topological modular forms*, and the group $\pi_{2n} \operatorname{tmf}$ the group of *topological modular forms of weight n*. The edge homomorphism of this spectral sequence is a homomorphism

$$\pi_{2n} \operatorname{tmf} \to M_n.$$

This map isn't quite surjective, and there is the following result of myself and Mark Mahowald

Proposition 4.6 *The image of the map $\pi_{2*} \operatorname{tmf} \to M_*$ has a basis given by the monomials*

$$a_{i,j,k}\, c_4^i c_6^j \Delta^k \qquad i, k \geq 0, j = 0, 1$$

where

$$a_{i,j,k} = \begin{cases} 1 & i > 0, j = 0 \\ 2 & j = 1 \\ 24/\gcd(24, k) & i, j = 0. \end{cases}$$

In the table below I have listed the first few homotopy groups of tmf

k	0	1	2	3	4	5	6	7	8
$\pi_k \operatorname{tmf}$	$\mathbb{Z}$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/24$	0	0	$\mathbb{Z}/2$	0	$\mathbb{Z} \oplus \mathbb{Z}/2$

9	10	11	12	13	14	15
$(\mathbb{Z}/2)^2$	$\mathbb{Z}/6$	0	$\mathbb{Z}$	$\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$

The homotopy homomorphism induced by the unit $S^0 \to \mathrm{tmf}$ of the ring tmf is the "tmf-degree," a ring homomorphism

$$\pi_* S^0 \to \pi_* \mathrm{tmf}\,.$$

The tmf-degree is an isomorphism in dimensions ≤ 6, and it take non-zero values on each of the classes represented by the Lie groups $SU(3)$, $U(3)$, $Sp(2)$, $Sp(1)\times Sp(2)$, G_2, $U(1)\times G_2$, regarded as framed manifolds via their left invariant framings. Thus, combined with the Hopf-invariant and the invariants coming from KO-theory, the tmf-degree accounts for all of $\pi_* S^0$ for $* \leq 15$. In fact the Hopf-invariant and the invariants coming from KO can also be described in terms of tmf and nearly all of $\pi_* S^0$ for $* < 60$ can be accounted for.

5. θ-series

5.1. Cohomology rings as rings of functions

Consider the computation

$$H^*(CP^\infty; \mathbb{Z}) = \mathbb{Z}[x].$$

On one hand this tells us something about the cell structure of complex projective space; the cohomology class x^n is "dual" to the cell in dimension $2n$. On the other hand, a polynomial is a function on the affine line, and the elements of $H^*(CP^\infty)$ tell us something about the affine line. Combining these, the prospect presents itself, of using the cell structure of one space to get information about the function theory of another.

We will apply this not to ordinary cohomology, but to the cohomology theory E_A. Before doing so, more of the function theoretic aspects of E_A need to be spelled out. By construction, the ring $E_A^0(CP^\infty)$ is the ring of functions on the formal completion of C at the point e at ∞. The ring $E_A^0(HP^\infty)$ is the ring of functions which are invariant under the involution

$$\begin{aligned} \tau(x) &= x \\ \tau(y) &= -y - a_1 x - a_3 \end{aligned} \tag{5.7}$$

given by the "inverse" in the group law. The map

$$E_A^0(CP^\infty) \to E_A^0(\mathrm{pt})$$

corresponds to evaluation at e, and the reduced cohomology group $\tilde{E}_A^0(CP^\infty)$ to the ideal of formal function vanishing at e. Note that this is consistent with the definition of A_2 as sections of the line bundle ω:

$$A_2 = \pi_2 E_A = \tilde{E}_A^0(S^2) = I/I^2 = H^0\left(\mathcal{O}_C(-e)/\mathcal{O}_C(-2e)\right).$$

Now $\tilde{E}^0_A(CP^\infty)$ is the cohomology group of the Thom complex $E^0_A(CP^{\infty L})$. More generally, there is an additive correspondence

$$\{\text{virtual representations of } U(1)\} \leftrightarrow \{\text{divisors on } C\},$$

under which a virtual representation V corresponds to a divisor D for which

$$E^0\left((CP^\infty)^V\right) = H^0\left(C^f; \prime(D) \otimes \left(\Omega^1\right)^{\otimes \mu(V)}\right),$$

where $\mu(V)$ is the multiplicity of the trivial representation in V. There is a similar correspondence between even functions with divisors of the form $D + \tau^* D$, virtual representations V of $SU(2)$ and $E^0_A\left((HP^\infty)^V\right)$ (with τ the involution (5.7)).

5.2. The Hopf fibration and the Weierstrass $\mathcal{P}$-function

Consider the function x in the Weierstrass equation (4.5). This function has a double pole at e. We now ask if there is a "best" x to choose, ie, a function with a double pole at e which is invariant under the action of r, s and t. Such a function will be an eigenvector for λ with eigenvalue λ^2. It is more convenient to search for an quantity which is invariant under λ as well, so instead we search for a quadratic differential on C, i.e. a section

$$x' \in H^0\left(C; \Omega^1(e)^{\otimes 2}\right)$$

which is invariant under r, s, t and λ. Now the space $H^0(C; \Omega^1(e)^2)$ has dimension 2, and sits in a short exact sequence of vector spaces over $\operatorname{Spec} A$

$$0 \to \omega^2 \to H^0(C; \Omega^1(e)^2) \to \mathcal{O}_A \to 0, \tag{5.8}$$

where ω is the line bundle of invariant differentials on C, and the second map is the "residue at e". This sequence is G-equivariant, and defines an element of

$$\operatorname{Ext}^1(\mathcal{O}, \omega^2) = H^1(G; A_4).$$

The obstruction to the existence of an x' with residue 1 is the Yoneda class ν of this extension. Completing the square and cube in (4.5), gives the G-invariant expression

$$\left(12x + a_1^2 + 4a_2\right) u^2,$$

so that $12\nu = 0$. In fact the group $H^1(G; A_4)$ is cyclic of order 12 with ν as a generator. The group $\pi_3 \operatorname{tmf} = \mathbb{Z}/24$ is assembled from $H^1(G; A_4)$ and $H^3(G; A_6) = Z/2$, and sits in an exact sequence

$$0 \to H^3(G; A_6) \to \pi_3 \operatorname{tmf} \to H^1(G; A_4) \to 0.$$

This 12 can also be seen transcendentally. Over the complex numbers, a choice of x is given by the Weierstrass $\mathcal{P}$-function:

$$\mathcal{P}(z, \tau) = \frac{1}{z^2} + \sum_{0 \neq (m,n) \in \mathbb{Z}^2} \frac{1}{(z - m\tau - n)^2} - \frac{1}{(m\tau + n)^2}.$$

The Fourier expansion of $\mathcal{P}(z,\tau)\,dz^2$ is (with $q = e^{2\pi i \tau}$ and $u = e^{2\pi i z}$)

$$\mathcal{P}(z,\tau)\,dz^2 = \left(\sum_{n\in\mathbb{Z}} \frac{q^n u}{(1-q^n u)^2} + \frac{1}{12} - 2\sum_{n\geq 1}\frac{q^n}{(1-q^n)^2}\right)\left(\frac{du}{u}\right)^2.$$

Note that all of the Fourier coefficients of $\mathcal{P}$ are integers, except for the constant term, which is $1/12$. This is the same 12.

Under the correspondence between divisors and Thom complexes, the differential x' corresponds to a G-invariant element

$$x'_{\text{top}} \in E_A^0\left(HP^{(2-V)}\right),$$

with V the defining representation of $SU(2)$. Now the spectrum HP^{2-V} has a (stable) cell decomposition

$$HP^{2(1-L)} = S^0 \cup_\nu e^4 \cup \ldots$$

with one cell in every real dimension $4k$. The 4-cell is attached to the 0-cell by the stable Hopf map $\nu : S^7 \to S^4$, which generates $\pi_3(S^0) = \mathbb{Z}/24$. The restriction of the quadratic differential x'_{top} to the zero cell is given by the residue at e, and the obstruction to the existence of an G-invariant x'_{top} with residue k is the image of k under the connecting homomorphism

$$H^0(G;A) \to H^1\left(G; E_A^0\left(HP^{2-V}/S^0\right)\right). \tag{5.9}$$

To evaluate (5.9), note that the map

$$H^1\left(G; E_A^0\left(HP^{2-V}/S^0\right)\right) \to H^1\left(G; E_A^0(S^4)\right) = H^1\left(G; A_4\right)) = \mathbb{Z}/12$$

is projection onto a summand, and the image of (5.9) is contained in this summand. Thus the obstruction to the existence of the quadratic differential x_{top} is the same $k \in \mathbb{Z}/12$. In this way, the theory of topological modular forms relates the Hopf map ν to the constant term in the Fourier expansion of the Weierstrass $\mathcal{P}$-function, and to the existence of a certain quadratic differential on the universal elliptic curve.

5.3. Lattices and their θ-series

There is a slightly more sophisticated application of these ideas to the theory of even unimodular lattices. Suppose that L is a positive definite, even unimodular lattice of dimension $2d$. The *theta function* of L, θ_L is the generating function

$$\begin{aligned}\theta_L(q) &= \sum_{\ell\in L} q^{\frac{1}{2}\langle \ell,\ell\rangle} \\ &= \sum_{n\geq 0} L_n q^n, \\ L_n &= \#\{\ell \mid \langle l, l\rangle = 2n\}.\end{aligned}$$

It follows from the Poisson summation formula that $\theta_L(q)$ is the q-expansion of a modular form over $\mathbb{Z}$ of weight d, and so lies in the ring

$$\mathbb{Z}[c_4, c_6, \Delta]/(c_4^3 - c_6^2 - 1728\Delta) \subset Z[\![q]\!],$$
$$c_4 = 1 + 240 \sum_{n>0} \sigma_3(n) q^n,$$
$$c_6 = 1 - 504 \sum_{n>0} \sigma_5(n) q^n,$$
$$\Delta = q \prod_{n=1}^{\infty} (1 - q^n)^{24}.$$

Since the group of modular forms of a given weight is finitely generated, the first few L_n determine the rest. This leads to many restrictions on the distributions of lengths of vectors in a positive definite, even unimodular lattice.

The θ-series of L is the value at $z = 0$ of the θ-function

$$\theta(z, \tau) = \sum_{\ell \in L} e^{\pi i \langle \ell, \ell \rangle \tau + 2\pi i \langle \ell, z \rangle}, \qquad z \in \mathbb{C} \otimes L, \quad q = e^{2\pi i \tau},$$

which, under the correspondence between divisors and representations has the following topological interpretation. Let V be any d-dimensional (complex) virtual representation of $U(1) \otimes L$ with the property that $c_1(V) = 0$ and $c_2(V)$ corresponds to the quadratic form, under the isomorphism

$$H^2\left(B(U(1) \otimes L); \mathbb{Z}\right) = \operatorname{Sym}^2 L^*.$$

For example, if $A = (a_{ij})$ is the matrix of the quadratic form with respect to some basis, then V could be taken to be

$$V = \mathbb{C}^d + \frac{1}{2} \sum_{i,j} a_{ij}(1 - L_i)(1 - L_j),$$

where L_i is the character of $U(1) \otimes L$ dual to the i^{th} basis element. The series $\theta(z, \tau)$ corresponds to a G-invariant element

$$\theta^{\text{top}} \in E_A^0((BU(1) \otimes L)^V).$$

The restriction of θ^{top} to $\{\text{pt}\}^V = S^{2n}$ is an element of $H^0(G; A_{2d})$, i.e. an algebraic modular form of weight d. This modular form is θ_L.

Now the Thom spectrum $(BU(1) \otimes L)^V$ has a stable cell decomposition

$$S^{2d} \cup \bigvee^{2d} e^{2d+2} \cup \bigvee^{d(2d+1)} e^{2d+4} \cup \cdots.$$

Since $c_1(V) = 0$, the cells of dimension $2d+2$ are not attached to the $(2d)$-cell. The assumptions on the quadratic form, and on $c_2(V)$ imply that one of the $(2d+4)$-cells is attached to the $(2d)$-cell by (a suspension of) the stable Hopf map ν. The presence of this attaching map implies the following mod 24 congruence on θ_L.

Theorem 5.10 *Suppose L is a positive definite, even unimodular lattice of dimension $24k$. Write*

$$\theta_L(q) = c_4^{3k} + x_1 c_4^{3(k-1)}\Delta + \cdots + x_k\Delta^k.$$

Then

$$x_k \equiv 0 \mod 24. \tag{5.11}$$

The above result was originally proved by Borcherds [12] as part of his investigation into infinite product expansions for automorphic forms on certain indefinite orthogonal groups. The above topological proof can be translated into the language of complex function theory. The details are in the next section.

The congruence of Theorem 5.10 together with Proposition 4.6 give the following

Proposition 5.12 *Suppose L is a positive definite, even unimodular lattice of dimension $2d$. There is an element $\theta_L^{top} \in \mathrm{tmf}^0(S^{2d})$ whose image in M_d is θ_L.*

It can also be shown that the G-invariant

$$\theta^{\mathrm{top}} \in H^0(G; E_A^0(B(U(1)\otimes L)^V))$$

is truly topological in the sense that it is the representative in the E_2-term of the spectral sequence

$$H^0(G; E_A^0(B(U(1)\otimes L)^V)) \Rightarrow \mathrm{tmf}^0(B(U(1)\otimes L)^V),$$

of an element in $\mathrm{tmf}^0(B(U(1)\otimes L)^V)$. I don't know of a direct construction of these truly topological theta series.

5.4. An analytic proof of Theorem 5.10

The analytic interpretation of the proof of Theorem 5.10 establishes the result in the form

$$\mathrm{Res}_{\Delta=0}\frac{\theta_L(q)}{\Delta^k}\frac{d\Delta}{\Delta} \equiv \mathrm{Res}_{q=0}\frac{\theta_L(q)}{\Delta^k}\frac{dq}{q} \equiv 0 \mod 24. \tag{5.13}$$

The equivalence of (5.11) with (5.13) follows easily from the facts

$$c_4 \equiv 0 \mod 24,$$
$$\frac{d\Delta}{\Delta} \equiv \frac{dq}{q} \mod 24.$$

Set $u = e^{2\pi i z}$, and for a vector $\mu \in L$ let

$$\phi_\mu(z,\tau) = \frac{\sum_{\ell\in L} q^{\frac{1}{2}\langle \ell,\ell\rangle} u^{\langle \mu,\ell\rangle}}{\sigma(q,u)^{\langle \mu,\mu\rangle}},$$

where

$$\sigma(q,u) = u^{\frac{1}{2}}(1-u^{-1})\prod_{n=1}^{\infty}\frac{(1-q^n u)(1-q^n u^{-1})}{(1-q^n)^2}$$

is the Weierstrass σ-function. It is immediate from the definition that

$$\phi_\mu(-z,\tau) = \phi_\mu(z,\tau),$$

and it follows from the modular transformation formula for θ that for

$$m,n \in \mathbb{Z}, \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(Z),$$

the function $\phi_\mu(z,\tau)$ satisfies

$$\phi_\mu(z + m\tau + n, \tau) = \phi_\mu(z,\tau), \tag{5.14}$$

$$\phi_\mu\left(\frac{z}{c\tau+d}, \frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^{d/2}\,\phi_\mu(z,\tau). \tag{5.15}$$

These identities are equivalent to saying that if we write, with $x = 2\pi i z$,

$$\phi_\mu(z,\tau) = x^{-\langle \mu,\mu\rangle}\left(\phi_\mu^{(0)} + \phi_\mu^{(2)}x^2 + \dots\right),$$

then $\phi_\mu^{(2k)}$ is a modular form of weight $d/2 + 2k$.

It is the term $\phi_\mu^{(2)}$ that stores the information about the $(2d+4)$-cells of the spectrum $B(U(1)\otimes L)^V$. A little computation shows that

$$\phi_\mu^{(2)} = \theta_\mu - \frac{\langle \mu,\mu\rangle}{24}\theta_L, \tag{5.16}$$

where

$$\theta_\mu = \sum_{\ell\in L} q^{\frac{1}{2}\langle \ell,\ell\rangle}\frac{\langle \ell,\mu\rangle^2}{2} - \langle \mu,\mu\rangle \sum_{n\geq 1}\frac{q^n}{(1-q^n)^2}.$$

Because of (5.15), the expression

$$\frac{\phi_\mu^{(2)}}{\Delta^k}\frac{dq}{q}$$

is a meromorphic differential on the projective j-line. It's only pole is at $q = 0$. It follows that[2]

$$\mathrm{Res}_{q=0}\frac{\phi_\mu^{(2)}}{\Delta^k}\frac{dq}{q} = 0,$$

and hence

$$\mathrm{Res}_{q=0}\frac{\theta_\mu}{\Delta^k}\frac{dq}{q} = \frac{\langle \mu,\mu\rangle}{24}\mathrm{Res}_{q=0}\frac{\theta_L}{\Delta^k}\frac{dq}{q}. \tag{5.17}$$

To deduce (5.13), define

$$p : L\otimes \mathbb{Z}/2 \to \mathbb{Z}/2$$

[2]I learned this trick from Borcherds, who told me he learned it from a referee.

by letting $p(\mu)$ be the value of (5.17) reduced modulo 2. The left hand side of (5.17) shows that p does in fact takes its values in $\mathbb{Z}/2$. Making use of the symmetry $\ell \mapsto -\ell$, it also shows that p is linear. The right hand side shows that p is quadratic with underlying bilinear form

$$p(\mu_1 + \mu_2) - p(\mu_1) - p(\mu_2) = \langle \mu_1, \mu_2 \rangle \frac{1}{12} \operatorname{Res}_{q=0} \frac{\theta_L}{\Delta^k} \frac{dq}{q}.$$

It follows that

$$\langle \mu_1, \mu_2 \rangle \frac{1}{12} \operatorname{Res}_{q=0} \frac{\theta_L}{\Delta^k} \frac{dq}{q} \equiv 0 \mod 2.$$

Since L is unimodular, there are vectors μ_1 and μ_2 with

$$\langle \mu_1, \mu_2 \rangle = 1$$

and so

$$\frac{1}{12} \operatorname{Res}_{q=0} \frac{\theta_L}{\Delta^k} \frac{dq}{q} \equiv 0 \mod 2.$$

This is (5.13).

6. Topological modular forms as cobordism invariants

6.1. The Atiyah-Bott-Shapiro map

As described in §2, the part of the homotopy groups of spheres that is best understood geometrically is the part that is captured by KO-theory. The geometric interpretations rely on the fact that all of the corresponding framed cobordism invariants can be expressed in terms of invariants of Spin-cobordism. From the point of view of topology what makes this possible is the factorization

$$S^0 \to M\mathrm{Spin} \to KO$$

of the unit in KO-theory, through the Atiyah-Bott-Shapiro map $M\mathrm{Spin} \to KO$. The Atiyah-Bott-Shapiro map is constructed using the representations of the spinor groups, and relies on knowing that for a space X, elements of $KO^0(X)$ are represented by vector bundles over X. It gives a KO-theory Thom isomorphism for Spin-vector bundles, and is topological expression for the index of the Dirac operator. Until recently it was not known how to produce this map by purely homotopy theoretic means.

The framed cobordism invariants coming from tmf cannot be expressed directly in terms of Spin-cobordism. One way of seeing this is to note that the group $\pi_3 M\mathrm{Spin}$ is zero. The fact that the map

$$\pi_3 S^0 \to \pi_3 \mathrm{tmf}$$

is an isomorphism prevents a factorization of the form

$$S^0 \to M\text{Spin} \to \text{tmf}\,.$$

In some sense this is all that goes wrong. Let $BO\langle 8\rangle$ be the 7-connected cover of $B\text{Spin}$, and $MO\langle 8\rangle$ the corresponding Thom spectrum. We will see below that there is a factorization of the unit

$$S^0 \to MO\langle 8\rangle \to \text{tmf}\,.$$

There is not yet a geometric interpretation of $\text{tmf}^0(X)$, so the construction of a map $MO\langle 8\rangle \to \text{tmf}$ must be made using homotopy theoretic methods. The key to doing this is to exploit the E_∞-ring structures on the spectra involved.

6.2. E_∞-maps

The spectra $MO\langle 8\rangle$, $M\text{Spin}$, tmf and KO are all E_∞-ring spectra, and from the point of view of homotopy theory it turns out easier to construct E_∞ maps

$$M\text{Spin} \to KO \quad \text{and} \quad MO\langle 8\rangle \to \text{tmf}$$

than merely to produce maps of spectra. In fact the homotopy type of the spaces of E_∞-maps

$$E_\infty\left(MO\langle 8\rangle, \text{tmf}\right) \quad \text{and} \quad E_\infty\left(M\text{Spin}, KO\right)$$

can be fairly easily identified using homotopy theoretic methods. In this section I will describe their sets of path components.

A map $\phi : M\text{Spin} \to KO$ determines a Hirzebruch genus with an even characteristic series $K_\phi(z) \in \mathbb{Q}[\![x]\!]$. Define the *characteristic sequence* of ϕ to be the sequence of rational numbers

$$(b_2, b_4, \cdots)$$

given by

$$\log\left(K_\phi(z)\right) = -2\sum_{n>0} b_n \frac{x^n}{n!}.$$

We use this sequence to form the *characteristic map*

$$\pi_0 E_\infty\left(M\text{Spin}, KO\right) \to \{(b_2, b_4, \cdots) \mid b_{2i} \in \mathbb{Q}\}$$

from the set of homotopy classes of E_∞ maps $M\text{Spin} \to KO$ to the set of sequences of rational numbers.

Let B_n denote the n^{th} Bernoulli number defined by

$$\frac{x}{e^x - 1} = \sum_{n\geq 0} B_n \frac{x^n}{n!}.$$

The following result is due to Matthew Ando, myself, and Charles Rezk:

Theorem 6.18 *The characteristic map*

$$\pi_0 E_\infty (M\mathrm{Spin}, KO) \to \{(b_2, b_4, \cdots) \mid b_{2i} \in \mathbb{Q}\}$$

gives an isomorphism of $\pi_0 E_\infty (M\mathrm{Spin}, KO)$ *with the set of sequences*

$$(b_2, b_4, b_6, \dots)$$

for which

i) $b_n \equiv B_n/2n \mod \mathbb{Z}$;

ii) *for each odd prime* p *and each* p*-adic unit* c,

$$m \equiv n \mod p^k(p-1)$$
$$\implies (1-c^n)(1-p^{n-1})b_n \equiv (1-c^m)(1-p^{m-1})b_m \mod p^{k+1};$$

iii) *for each* 2*-adic unit* c

$$m \equiv n \mod 2^k$$
$$\implies (1-c^n)(1-2^{n-1})b_n \equiv (1-c^m)(1-2^{m-1})b_m \mod 2^{k+2}.$$

Remark 6.19 By the Kummer congruences, the sequence with $b_{2n} = B_{2n}/(4n)$ comes from an E_∞-map $M\mathrm{Spin} \to KO$. The associated characteristic series is

$$\frac{x}{e^{x/2} - e^{-x/2}} = \frac{x/2}{\sinh(x/2)}$$

and the underlying map of spectra coincides with the one constructed by Atiyah-Bott-Shapiro. Theorem 6.18 therefore gives a purely homotopy theoretic construction of this map.

We now turn to describing the set of homotopy classes of E_∞ maps

$$MO\langle 8\rangle \to \mathrm{tmf}\,.$$

Associated to a multiplicative map $MO\langle 8\rangle \to \mathrm{tmf}$ is a characteristic series of the form

$$K_\phi(z) = \sum a_{2n} z^{2n} \qquad a_{2n} \in \mathbb{Q}[\![q]\!]$$

which is well-defined up to multiplication by the exponential of a quadratic function in z. We associate to such a series, a sequence

$$(g_4, g_6 \cdots) \quad g_{2n} \in \mathbb{Q}[\![q]\!] \tag{6.20}$$

according to the rule

$$\log(K_\phi(z)) = -2 \sum_{n>0} g_n \frac{x^n}{n!}.$$

For $n > 2$ the terms g_n is independent of the quadratic exponential factor, and is the q-expansion of a modular form of weight n (see [27, 6]). This defines the *characteristic map*

$$\pi_0 E_\infty (MO\langle 8\rangle, \mathrm{tmf}) \to \{(g_2, g_4, \cdots) \mid g_{2n} \in M_{2n} \otimes \mathbb{Q}\} .$$

Following Serre [52] let G_{2k} denote the (un-normalized) Eisenstein series of weight $2k$

$$G_{2k} = -\frac{B_{2k}}{4k} + \sum_{n>0} \sigma_{2k-1}(n) q^n, \qquad \sigma_{2k-1}(n) = \sum_{d|n} d^{2k-1},$$

and for a prime p, let G^*_{2k} be the (un-normalized) p-adic Eisenstein series

$$G^*_{2k} = -(1 - p^{2k-1})\frac{B_{2k}}{4k} + \sum_{n>0} \sigma^*_{2k-1}(n) q^n, \qquad \sigma^*_{2k-1}(n) = \sum_{\substack{d|n \\ (d,p)=1}} d^{2k-1}.$$

We will also need the Atkin (U) and Verschiebung (V) operators on p-adic modular forms of weight k (See [52, §2.1]). For a p-adic modular form

$$f = \sum_{n=0}^{\infty} a_n q^n$$

of weight k, one defines

$$f|U = \sum_{n=0}^{\infty} a_{pn} q^n \qquad f|V = \sum_{n=0}^{\infty} a_n q^{pn}.$$

Finally, we set

$$f^* = f - p^{k-1} f|V.$$

(This gives two meanings to the symbol G^*_{2k}, which are easily checked to coincide.)

Proposition 6.21 *The image of the characteristic map*

$$\pi_0 E_\infty (MO\langle 8\rangle, \mathrm{tmf}) \to \{(g_2, g_4, \cdots) \mid g_{2n} \in M_{2n} \otimes \mathbb{Q}\}$$

is the set of sequences (g_{2n}) satisfying

i) $g_{2n} \equiv G_{2n} \mod M_{2n}$

ii) *For each odd prime p and each p-adic unit c,*

$$m \equiv n \mod p^k(p-1) \implies (1 - c^n) g^*_n \equiv (1 - c^m) g^*_m \mod p^{k+1},$$

iii) *For each 2-adic unit c,*

$$m \equiv n \mod 2^k \implies (1 - c^n) g^*_n \equiv (1 - c^m) g^*_m \mod 2^{k+2}.$$

iv) *For each prime p, $g_m^*|U = g_m^*$.*

The characteristic map is a principle A-bundle over its image, where A is a group isomorphic to a countably infinite product of $\mathbb{Z}/2$'s.

Remark 6.22 The group A occurring in the final assertion of Proposition 6.21 can be described explicitly in terms of modular forms. The description is somewhat technical, and has been omitted.

By the Kummer congruences, the sequence of Eisenstein series $(G_4, G_6, \cdots)$ satisfy the conditions of Proposition 6.21. The corresponding characteristic series can be taken to be

$$\frac{z/2}{\sinh(z/2)} \prod_{n \geq 1} \frac{(1-q^n)^2}{(1-q^n e^z)(1-q^n e^{-z})},$$

and the genus is the Witten genus ϕ_W [57, 49, 21]. This gives the following corollary, which was the main conjecture of [?].

Corollary 6.23 *There is an E_∞-map $MO\langle 8\rangle \to$ tmf whose underlying genus is the Witten genus.*

Remark 6.24 This refined Witten genus is only specified up to action by an element of the group A occurring in Proposition 6.21. It looks as if a more careful analysis could lead to specifying a single E_∞-map $MO\langle 8\rangle \to$ tmf, but this has not yet been carried out.

6.3. The image of the cobordism invariant

The existence of a refined Witten genus has an application to the theory of even unimodular lattices. The following two results are due to myself and Mark Mahowald.

Theorem 6.25 *Let $MO\langle 8\rangle \to$ tmf be any multiplicative map whose underlying genus is the Witten genus. Then the induced map of homotopy groups $\pi_* MO\langle 8\rangle \to \pi_*$ tmf is surjective.*

Combining this with Proposition 5.12 then gives

Corollary 6.26 *Let L be a positive definite, even unimodular lattice of dimension $2d$. There exists a 7-connected manifold M_L of dimension $2d$, whose Witten genus is the θ-function of L, i.e.*

$$\phi_W(M) = \theta_L.$$

In case L is the Leech lattice, the existence of M_L gives an affirmative answer to Hirzebruch's "Prize Question" [21].

6.4. Spectra of units and E_∞-maps

The proofs of Theorems 6.18 and 6.21 come down to understanding the structure of the group of units in $KO^0(X)$ and $\mathrm{tmf}^0(X)$.

For a ring spectrum R, let $Gl_1(R)$ be the classifying space for the group of units in R-cohomology:

$$[X, Gl_1(R)] = R^0(X)^\times.$$

If $R = \{R_n, t_n\}$, then $Gl_1(R)$ is part of the homotopy pullback square

$$\begin{array}{ccc} Gl_1(R) & \longrightarrow & R_0 \\ \downarrow & & \downarrow \\ \pi_0 R^\times & \longrightarrow & \pi_0 R. \end{array} \tag{6.27}$$

When R is an A_∞-ring spectrum, $Gl_1(R)$ has a classifying space $BGl_1(R)$. When R is E_∞, then $Gl_1(R)$ is an infinite loop space, i.e. there is a spectrum $gl_1(R)$ with $gl_1(R)_0 = Gl_1(R)$ (and $gl_1(R)_1 = BGl_1(R)$). The space $BGl_1(S^0)$ is the classifying space for unoriented stable spherical fibrations, and the map

$$BO\langle 8\rangle \to BGl_1(S^0) \tag{6.28}$$

whose associated Thom spectrum is $MO\langle 8\rangle$, is an infinite loop map. More specifically, let $bo\langle 8\rangle$ be the 7-connective cover of the spectrum KO. Then

$$\left(\Sigma^{-1} bo\langle 8\rangle\right)_1 = BO\langle 8\rangle,$$

and there is a map of spectra

$$\Sigma^7 bo\langle 8\rangle \to gl_1(S^0)$$

for which the induced map $\left(\Sigma^{-1} bo\langle 8\rangle\right)_1 \to \left(gl_1(S^0)\right)_1$ becomes (6.28).

The following result is what makes it easier to construct E_∞-maps than merely maps of spectra.

Proposition 6.29 *The space $E_\infty(M\mathrm{Spin}, KO)$ is canonically homotopy equivalent to the space of factorizations*

$$\begin{array}{ccc} gl_1(S^0) & \longrightarrow & gl_1(S^0) \cup C\Sigma^{-1} bo\langle 4\rangle \\ \downarrow & \swarrow & \\ gl_1(KO), & & \end{array}$$

and the space $E_\infty(MO\langle 8\rangle, \mathrm{tmf})$ is canonically homotopy equivalent to the space of factorizations

$$\begin{array}{ccc} gl_1(S^0) & \longrightarrow & gl_1(S^0) \cup C\Sigma^{-1} bo\langle 8\rangle \\ \downarrow & \swarrow & \\ gl_1(\mathrm{tmf}). & & \end{array}$$

6.5. The Atkin operator and the spectrum of units in tmf

Proposition 6.29 emphasizes the important role played by the spectrum of units $gl_1(\mathrm{tmf})$ and $gl_1(KO)$. Getting at the homotopy type of these spectra uses work of Bousfield [13] and Kuhn [31]. Fix a prime p let $K(n)$ denote the n^{th} Morava K-theory at p, and

$$L_{K(n)} : \text{Spectra} \to \text{Spectra}$$

the localization functor with respect to $K(n)$ (see [46, 48]). Bousfield (in case $n = 1$) and Kuhn (in case $n > 1$) construct a functor

$$\Sigma^{K(n)} : \text{Spaces} \to \text{Spectra}$$

and a natural equivalence of $\Sigma^{K(n)}(E_0)$ with $L_{K(n)}E$. The spectrum $\Sigma^{K(n)}(X)$ depends only on the connected component of X containing the basepoint. In the special case of an E_∞-ring spectrum E it gives (because of (6.27)) a canonical equivalence

$$L_{K(n)}\, gl_1\, E \approx L_{K(n)}E,$$

which, when composed with the localization map $gl_1\, E \to L_{K(n)}\, gl_1\, E$, leads to a "logarithm"

$$\log^E_{K(n)} : gl_1\, E \to L_{K(n)}E.$$

Bousfield showed that for KO, the logarithm

$$\log^{KO}_{K(1)} : gl_1\, KO \to L_{K(1)}KO = KO_p$$

becomes an equivalence after completing at p and passing to 2-connected covers. Charles Rezk has recently shown that for a space X, the map

$$\log^{KO}_{K(1)} : KO^0(X)^\times \to KO^0_p(X)$$

is given by the formula

$$\frac{1}{p}\log\left(\frac{\psi_p(x)}{x^p}\right).$$

This equivalence between the multiplicative and additive groups of K-theory was originally observed by Sullivan, and proved by Atiyah-Segal [9].

In the case of tmf, Paul Goerss, Charles Rezk and I have shown that the Bousfield-Kuhn logarithms leads to a commutative diagram

$$\begin{array}{ccc} gl_1\,\mathrm{tmf} & \xrightarrow{\log_p^{\mathrm{tmf}}} & \mathrm{tmf}_p \\ \downarrow & & \downarrow \\ L_{K(1)}\,\mathrm{tmf} & \xrightarrow[1-U]{} & L_{K(1)}\,\mathrm{tmf} \end{array} \tag{6.30}$$

which becomes homotopy Cartesian after completing at p and passing to 3-connected covers. The spectrum tmf_p is the p-adic completion of tmf, and the map $\log_p^{\mathrm{tmf}}$ has a description in terms of modular forms, but to describe it would take us outside

the scope of this paper. The spectrum $L_{K(1)}\,\mathrm{tmf}$ is the topological analogue of the theory of p-adic modular forms of Serre [52] and Katz [30], and the map U is the topological Atkin operator. One noteworthy feature of this square is that it locates the Atkin operator in the theory of all modular forms (and not just p-adic modular forms). Another is that it connects the failure of $\log_p^{\mathrm{tmf}}$ to be an isomorphism with the spectrum of U. For example, let F denote the fiber of

$$gl_1\,\mathrm{tmf}_p \xrightarrow{\log_p^{\mathrm{tmf}}} \mathrm{tmf}_p\,.$$

By the square (6.30)

$$\pi_{23}F = \begin{cases} \mathbb{Z}_p & p = 691 \\ \mathbb{Z}_p \oplus \mathbb{Z}_p/(\tau(p) - (p^{11}+1)) & p \neq 691, \end{cases}$$

where τ is the Ramanujan τ function, defined by

$$q\prod_{n=1}^{\infty}(1-q^n)^{24} = \sum_{k=1}^{\infty}\tau(k)q^k.$$

For $p \neq 691$ the torsion subgroup of $\pi_{23}F$ has order determined by the p-adic valuation of $(\tau(p)-1)$. The only primes less that $35,000$ for which $\tau(p) \equiv 1 \mod p$ are 11, 23, and 691. It is not known whether or not $\tau(p) \equiv 1 \mod p$ holds for infinitely many primes.

The fiber of $\log_p^{\mathrm{tmf}}$ seems to store quite a bit of information about the spectrum of the Atkin operator and p-adic properties of modular forms. Investigating its homotopy type looks like interesting prospect for algebraic topology.

References

[1] J. F. Adams, *On the structure and applications of the Steenrod algebra*, Commentarii Mathematici Helvetici **32** (1958), 180–214.

[2] ———, *On the groups J(X), IV*, Topology **5** (1966), 21–71.

[3] J. F. Adams, *Lectures on generalised cohomology*, Category Theory, Homology Theory and their Applications, III (Battelle Institute Conference, Seattle, Wash., 1968, Vol. Three), Springer, Berlin, 1969, 1–138. MR 40 #4943

[4] J. F. Adams, *Stable homotopy and generalised homology*, University of Chicago Press, Chicago, 1974.

[5] J. Frank Adams, *Stable homotopy theory*, Springer-Verlag, Berlin, 1969. MR 40 #2065

[6] M. Ando, M. J. Hopkins, and N. P. Strickland, *Elliptic spectra, the Witten genus and the theorem of the cube*, Invent. Math. **146** (2001), no. 3, 595–687. MR 2002g:55009

[7] M. F. Atiyah, V. K. Patodi, and I. M. Singer, *Spectral asymmetry and Riemannian geometry. II*, Math. Proc. Cambridge Philos. Soc. **78** (1975), no. 3, 405–432. MR 53 #1655b

[8] ———, *Spectral asymmetry and Riemannian geometry. III*, Math. Proc. Cambridge Philos. Soc. **79** (1976), no. 1, 71–99. MR 53 #1655c

[9] M. F. Atiyah and G. B. Segal, *Exponential isomorphisms for λ-rings*, Quart. J. Math. Oxford Ser. (2) **22** (1971), 371–378. MR 45 #344

[10] M. F. Atiyah and I. M. Singer, *Index theory for skew-adjoint Fredholm operators*, Inst. Hautes Études Sci. Publ. Math. (1969), no. 37, 5–26. MR 44 #2257

[11] M. F. Atiyah and I. M. Singer, *The index of elliptic operators: V*, Annals of Mathematics **93** (1971), 139–149.

[12] Richard E. Borcherds, *Automorphic forms on* $\mathrm{o}_{s+2,2}(\mathbf{r})$ *and infinite products*, Invent. Math. **120** (1995), no. 1, 161–213. MR 96j:11067

[13] A. K. Bousfield, *Uniqueness of infinite deloopings for K-theoretic spaces*, Pacific J. Math. **129** (1987), no. 1, 1–31. MR 89g:55017

[14] W. Browder, *The Kervaire invariant of framed manifolds and its generalization*, Annals of Mathematics **90** (1969), 157–186.

[15] Henri Cartan and Jean-Pierre Serre, *Espaces fibrés et groupes d'homotopie. I. Constructions générales*, C. R. Acad. Sci. Paris **234** (1952), 288–290. MR 13,675a

[16] ———, *Espaces fibrés et groupes d'homotopie. II. Applications*, C. R. Acad. Sci. Paris **234** (1952), 393–395. MR 13,675b

[17] P. Deligne, *Courbes elliptiques: formulaire (d'après J. Tate)*, Modular Functions of One Variable III, Lecture Notes in Mathematics, vol. 476, Springer-Verlag, 1975, 53–73.

[18] Ethan S. Devinatz, Michael J. Hopkins, and Jeffrey H. Smith, *Nilpotence and stable homotopy theory. I*, Ann. of Math. (2) **128** (1988), no. 2, 207–241. MR 89m:55009

[19] A. D. Elmendorf, I. Kříž, M. A. Mandell, and J. P. May, *Modern foundations for stable homotopy theory*, Handbook of algebraic topology, North-Holland, Amsterdam, 1995, 213–253. MR 97d:55016

[20] A. D. Elmendorf, I. Kriz, M. A. Mandell, and J. P. May, *Rings, modules, and algebras in stable homotopy theory*, American Mathematical Society, Providence, RI, 1997, With an appendix by M. Cole. MR 97h:55006

[21] F. Hirzebruch, T. Berger, and R. Jung, *Manifolds and modular forms*, Aspects of Mathematics, vol. E 20, Friedr. Vieweg & Sohn, Braunschweig, 1994.

[22] M. J. Hopkins, *Global methods in homotopy theory*, Proceedings of the 1985 LMS Symposium on Homotopy Theory (J. D. S. Jones and E. Rees, eds.), 1987, 73–96.

[23] M. J. Hopkins and P. Goerss, *Multiplicative stable homotopy theory*, book in preparation.

[24] M. J. Hopkins, M. Mahowald, and H. R. Miller, *Elliptic curves and stable homotopy theory I*, in preparation.

[25] M. J. Hopkins, J. H. Palmieri, and J. H. Smith, *Vanishing lines in generalized Adams spectral sequences are generic*, Geom. Topol. **3** (1999), 155–165 (electronic). MR 2000c:55017

[26] M. J. Hopkins and I. M. Singer, *Quadratic functions in geometry, topology, and*

M-theory, math.AT/0211216.
[27] Michael J. Hopkins, *Topological modular forms, the Witten genus, and the theorem of the cube*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994) (Basel), Birkhäuser, 1995, 554–565. MR 97i:11043
[28] Michael J. Hopkins and Jeffrey H. Smith, *Nilpotence and stable homotopy theory. II*, Ann. of Math. (2) **148** (1998), no. 1, 1–49. MR 99h:55009
[29] Mark Hovey, Brooke Shipley, and Jeff Smith, *Symmetric spectra*, J. Amer. Math. Soc. **13** (2000), no. 1, 149–208. MR 2000h:55016
[30] Nicholas M. Katz, *p-adic properties of modular schemes and modular forms*, Modular functions of one variable, III (Proc. Internat. Summer School, Univ. Antwerp, Antwerp, 1972), Springer, Berlin, 1973, 69–190. Lecture Notes in Mathematics, Vol. 350. MR 56 #5434
[31] Nicholas J. Kuhn, *Morava K-theories and infinite loop spaces*, Algebraic topology (Arcata, CA, 1986), Springer, Berlin, 1989, 243–257. MR 90d:55014
[32] Gerd Laures, *On cobordism of manifolds with corners*, Trans. Amer. Math. Soc. **352** (2000), no. 12, 5667–5688 (electronic). MR 1 781 277
[33] W. Lellman and M. E. Mahowald, *The bo-adams spectral sequence*, Transactions of the American Mathematical Society **300** (1987), 593–623.
[34] L. G. Lewis, J. P. May, and M. Steinberger, *Equivariant stable homotopy theory*, Lecture Notes in Mathematics, vol. 1213, Springer-Verlag, New York, 1986.
[35] M. E. Mahowald, *The order of the image of the J-homomorphism*, Bulletin of the American Mathematical Society **76** (1970), 1310–1313.
[36] ———, *bo-resolutions*, Pacific Journal of Mathematics **192** (1981), 365–383.
[37] ———, *The image of J in the EHP sequence*, Annals of Mathematics **116** (1982), 65–112.
[38] J. Peter May, *E_∞ ring spaces and E_∞ ring spectra*, Springer-Verlag, Berlin, 1977, With contributions by Frank Quinn, Nigel Ray, and Jørgen Tornehave, Lecture Notes in Mathematics, Vol. 577. MR 58 #13008
[39] J. W. Milnor, *On the cobordism ring Ω^* and a complex analogue, Part I*, American Journal of Mathematics **82** (1960), 505–521.
[40] S. P. Novikov, *Methods of algebraic topology from the point of view of cobordism theory*, Izv. Akad. Nauk SSSR Ser. Mat. **31** (1967), 855–951. MR 36 #4561
[41] ———, *Rings of operations and spectral sequences of Adams type in extraordinary cohomology theories. U-cobordism and K-theory*, Dokl. Akad. Nauk SSSR **172** (1967), 33–36. MR 36 #884
[42] L. S. Pontryagin, *A classification of continuous transformations of a complex into a sphere. 1*, Doklady Akad. Nauk SSSR (N.S.) **19** (1938), 361–363.
[43] ———, *A classification of continuous transformations of a complex into a sphere. 2*, Doklady Akad. Nauk SSSR (N.S.) **19** (1938), 147–149.
[44] ———, *Homotopy classification of mappings of an $(n+2)$-dimensional sphere on an n-dimensional one*, Doklady Akad. Nauk SSSR (N.S.) **19** (1950), 957–959, Russian.
[45] D. G. Quillen, *On the formal group laws of unoriented and complex cobordism theory*, Bulletin of the American Mathematical Society **75** (1969), 1293–1298.
[46] D. C. Ravenel, *Localization with respect to certain periodic homology theories*,

American Journal of Mathematics **106** (1984), 351–414.

[47] ———, *Complex cobordism and stable homotopy groups of spheres*, Academic Press, New York, 1986.

[48] Douglas C. Ravenel, *Nilpotence and periodicity in stable homotopy theory*, Princeton University Press, Princeton, NJ, 1992, Appendix C by Jeff Smith. MR 94b:55015

[49] G. Segal, *Elliptic cohomology*, Séminaire Bourbaki 1987/88, Astérisque, vol. 161-162, Societe Mathematique de France, Féverier 1988, 187–201.

[50] Jean-Pierre Serre, *Sur la suspension de Freudenthal*, C. R. Acad. Sci. Paris **234** (1952), 1340–1342. MR 13,675d

[51] ———, *Sur les groupes d'Eilenberg-MacLane*, C. R. Acad. Sci. Paris **234** (1952), 1243–1245. MR 13,675c

[52] ———, *Formes modulaires et fonctions zêta p-adiques*, Modular functions of one variable, III (Proc. Internat. Summer School, Univ. Antwerp, 1972), Springer, Berlin, 1973, 191–268. Lecture Notes in Math., Vol. 350. MR 53 #7949a

[53] R. E. Stong, *Notes on cobordism theory*, Princeton University Press, Princeton, 1968.

[54] R. Thom, *Quelques propriétés globales des variétés differentiables*, Commentarii Mathematici Helvetici **28** (1954), 17–86.

[55] Friedhelm Waldhausen, *Algebraic K-theory of topological spaces. I*, Algebraic and geometric topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, Calif., 1976), Part 1, Amer. Math. Soc., Providence, R.I., 1978, 35–60. MR 81i:18014a

[56] ———, *Algebraic K-theory of topological spaces. II*, Algebraic topology, Aarhus 1978 (Proc. Sympos., Univ. Aarhus, Aarhus, 1978), Springer, Berlin, 1979, 356–394. MR 81i:18014b

[57] E. Witten, *The index of the Dirac operator in loop space*, Elliptic Curves and Modular Forms in Algebraic Topology (New York) (P. S. Landweber, ed.), Lecture Notes in Mathematics, vol. 1326, Springer-Verlag, 1988, 161–181.

[58] Edward Witten, *Five-brane effective action in M-theory*, J. Geom. Phys. **22** (1997), no. 2, 103–133, hep-th/9610234.

ICM 2002 · Vol. I · 319–344

Classification of Supersymmetries

Victor G. Kac*

Abstract

In the first part of my talk I will explain a solution to the extension of Lie's problem on classification of "local continuous transformation groups of a finite-dimensional manifold" to the case of supermanifolds. (More precisely, the problem is to classify simple linearly compact Lie superalgebras, i.e. topological Lie superalgebras whose underlying space is a topological product of finite-dimensional vector spaces). In the second part I will explain how this result is used in a classification of superconformal algebras. The list consists of affine superalgebras and certain super extensions of the Virasoro algebra. In the third part I will discuss representation theory of affine superalgebras and its relation to "almost" modular forms. Furthermore, I will explain how the quantum reduction of these representations leads to a unified representation theory of super extensions of the Virasoro algebra. In the forth part I will discuss representation theory of exceptional simple infinite-dimensional linearly compact Lie superalgebras and will speculate on its relation to the Standard Model.

Introduction

The theory of Lie groups and Lie algebras began with the 1880 paper [L] of S. Lie where he posed the problem of classification of "local continuous transformation groups of a finite-dimensional manifold" M and gave a solution to this problem when $\dim M = 1$ and 2.

The most important part of Lie's problem is the classification of the corresponding Lie algebras of vector fields on M up to "formal" isomorphism. A more invariant (independent of M) formulation is to classify linearly compact Lie algebras, i.e., topological Lie algebras whose underlying space is a topological product of discretely topologized finite-dimensional vector spaces [GS], [G2]. (Of course, it is well-known that it is impossible to classify even all finite-dimensional Lie algebras. What is usually meant be a "classification" is a complete list of simple algebras (no non-trivial ideas) and a description of semisimple algebras (no abelian ideals) in terms of simple ones.)

*Department of Mathematics, MIT, Cambridge, MA 02139, USA. E-mail: kac@math.mit.edu

It turned out that a solution to this problem requires quite different methods in the cases of finite-dimensional and infinite-dimensional groups. The most important advance in the finite-dimensional case was made by W. Killing and E. Cartan at the end of the 19^{th} century who gave the celebrated classification of simple finite-dimensional Lie algebras over $\mathbb{C}$. The infinite-dimensional case was studied by Cartan in a series of papers written in the beginning of the 20^{th} century, which culminated in his classification of infinite-dimensional "primitive' Lie algebras of vector fields on a finite-dimensional manifold [C].

The advent of supersymmetry in theoretical physics in the 1970s motivated work on the "super" extension of Lie's problem. In the finite-dimensional case the latter problem was settled in [K2]. However, it took another 20 years before the problem was solved in the infinite-dimensional case [K7], [CK2], [CK3].

In the first part of my talk I will explain the classification of simple linearly compact Lie superalgebras. Remarkably, unlike in the Lie algebra case, the approach, based on the ideas of the papers [GS], [W], [K1] and [G2], is very similar in the finite- and infinite-dimensional cases.

The advent of conformal field theory in the mid-1980s motivated the work on classification and representation theory of superconformal algebras. In the second part of my talk I will explain how the classification of infinite-dimensional simple linearly compact Lie algebras is applied to classification of "linear" simple superconformal algebras. A complete list consists of the affine superalgebras, and of several series and one exceptional example of super extensions of the Virasoro algebra [FK]. (The most famous of these super extensions is the $N = 2$ superconformal algebra, which plays a fundamental role in the mirror symmetry theory.)

In the third part of my talk I will discuss representation theory of affine superalgebras [KW3], [KW4]. The key property of "admissible" representations of affine algebras is that their characters are modular functions. This is not so in the super case—for some mysterious reason, modular functions get replaced by closely related but more general functions, like Appell's function [KW4].

Next, I will explain how the quantum reduction of "admissible" representations of affine superalgebras leads to a unified representation theory of (not necessarily linear) super extensions of the Virasoro algebra [KRW],[KW5]. This gives rise to a large class of supersymmetric rational conformal field theories.

In the last part of my talk I will discuss representation theory of exceptional infinite-dimensional simple linearly compact Lie superalgebras [KR1]–[KR4]. I am convinced that this theory may have applications to "real" physics. The main reason for this belief is the exceptional Lie superalgebra $E(3|6)$ whose maximal compact group of automorphisms is the gauge group of the Standard Model (= a quotient of $SU_3 \times SU_2 \times U_1$ by a cyclic group of order 6). Furthermore, representation theory of $E(3|6)$ accurately predicts the number of generations of leptons (= 3), but not so accurately the number of generations of quarks (= 5) [KR2]. It is also striking that the inclusion of the gauge group of the Standard Model in SU_5, which is the gauge group of the Grand Unified Model, extends to the inclusion of $E(3|6)$ in $E(5|10)$, the largest exceptional linearly compact Lie superalgebra.

1. Classification of simple linearly compact Lie superalgebras.

1.1. First, recall some basic superalgebra terminology. A *superalgebra* is simply a $\mathbb{Z}/2\mathbb{Z}$-graded algebra:

$$S = S_{\overline{0}} + S_{\overline{1}}, \text{ where } S_\alpha S_\beta \subset S_{\alpha+\beta}, \alpha, \beta \in \mathbb{Z}/2\mathbb{Z} = \{\overline{0}, \overline{1}\}.$$

If $a \in S_\alpha$, one says that the *parity* $p(a)$ is equal to α. A *derivation* D of parity $p(D)$ of a superalgebra S is a vector space endomorphism satisfying condition

$$D(ab) = (Da)b + (-1)^{p(D)p(a)}a(Db).$$

The sum Der S of the spaces of derivations of parity $\overline{0}$ and $\overline{1}$ is closed under the super bracket:

$$[D, D_1] = DD_1 - (-1)^{p(D)p(D_1)}D_1D.$$

This super bracket satisfies super analogs of anticommutativity and Jacobi identity, hence defines what is called a *Lie superalgebra.* (The super anticommutativity axiom is $[a, b] = -(-1)^{p(a)p(b)}[b, a]$, and the super Jacobi identity axiom means that the operator $(\operatorname{ad} a)b := [a, b]$ is a derivation.)

One of the basic constructions is the *superization* which basically amounts to adding anticommuting indeterminates. In other words, given an algebra (associative or Lie) $\mathcal{A}$ we consider the Grassmann algebra $\mathcal{A}\langle n\rangle$ in n anticommuting indeterminates $\xi_1, \ldots, \xi_n$ over $\mathcal{A}$. This algebra carries a canonical $\mathbb{Z}/2\mathbb{Z}$-gradation defined by letting $p(\mathcal{A}) = \overline{0}, \quad p(\xi_i) = \overline{1}$. If $\mathcal{O}_m$ denotes the algebra of formal power series over $\mathbb{C}$ in m indeterminates, then $\mathcal{O}_m\langle n\rangle$ is the algebra over $\mathbb{C}$ of formal power series in m commuting indeterminates $x = (x_1, \ldots, x_m)$ and n anticommuting indeterminates $\xi = (\xi_1, \ldots, \xi_n)$: $x_ix_j = x_jx_i, \quad x_i\xi_j = \xi_jx_i, \quad \xi_i\xi_j = -\xi_j\xi_i$.

Note that the associative superalgebra $\mathcal{O}_m\langle n\rangle$ is linearly compact with respect to the topology for which the powers of the augmentation ideal $(x_1, \ldots, x_m, \xi_1, \ldots, \xi_n)$ form a fundamental system of neighborhoods of 0. The algebra $\mathcal{A}\langle n\rangle$ has odd (i.e., of parity $\overline{1}$) derivations $\partial/\partial\xi_i$ defined by

$$\frac{\partial}{\partial\xi_i}(a) = 0 \text{ for } a \in \mathcal{A}, \quad \frac{\partial}{\partial\xi_i}(\xi_j) = \delta_{ij},$$

and these derivations anticommute, i.e., $[\partial/\partial\xi_i, \partial/\partial\xi_j] = 0$.

The first basic example of a linearly compact Lie superalgebra is the Lie superalgebra denoted by $W(m|n)$, of all continuous derivations of the topological superalgebra $\mathcal{O}_m\langle n\rangle$:

$$W(m|n) = \left\{ \sum_{i=1}^{m} P_i(x,\xi)\frac{\partial}{\partial x_i} + \sum_{j=1}^{n} Q_j(x,\xi)\frac{\partial}{\partial\xi_j} \right\},$$

where $P_i(x,\xi), Q_j(x,\xi) \in \mathcal{O}_m\langle n\rangle$. In a more geometric language, this is the Lie superalgebra of all formal vector fields on a supermanifold of dimension $(m|n)$.

1.2. Cartan's theorem [C] states that a complete list of infinite-dimensional linearly compact simple Lie algebras over $\mathbb{C}$ consists of four series: the Lie algebra $W_m(=W(m|0))$ of all formal vector fields on an m-dimensional manifold, and its subalgebras S_m of divergenceless vector fields ($m>1$), H_m of Hamiltonian vector fields (m even), K_m of contact vector fields (m odd).

There is a unique way to extend *divergence* from W_m to $W(m|n)$ such that the divergenceless vector fields form a subalgebra:

$$\operatorname{div}\left(\sum_i P_i\frac{\partial}{\partial x_i}+\sum_j Q_j\frac{\partial}{\partial\xi_j}\right)=\sum_i\frac{\partial P_i}{\partial x_i}+\sum_j(-1)^{p(Q_j)}\frac{\partial Q_j}{\partial\xi_j}\,,$$

and the super analog of S_m is

$$S(m|n)=\{X\in W(m|n)|\operatorname{div}X=0\}\,.$$

In order to define super analogs of the Hamiltonian and contact Lie algebras H_m and K_m, introduce a super analog of the algebra of differential forms [K2]. This is an associative superalgebra over $\mathcal{O}_m\langle n\rangle$, denoted by $\Omega(m|n)$, on generators $dx_1,\dots,dx_m$, $d\xi_1,\dots,d\xi_n$ and defining relations: $dx_idx_j=-dx_jdx_i$, $dx_id\xi_j=d\xi_jdx_i$, $d\xi_id\xi_j=d\xi_jd\xi_i$, and the $\mathbb{Z}/2\mathbb{Z}$-gradation defined by: $p(x_i)=p(d\xi_j)=\bar{0}$, $p(\xi_j)=p(dx_i)=\bar{1}$. This superalgebra is linearly compact in the topology defined by powers of the augmentation ideal. The topological superalgebra $\Omega(m|n)$ carries a unique continuous derivation d of parity $\bar{1}$ such that $d(x_i)=dx_i$, $d(\xi_j)=d\xi_j$, $d(dx_i)=0$, $d(d\xi_j)=0$. The operator d has all the usual properties, e.g.: $df=\sum_i dx_i\frac{\partial f}{\partial x_i}+\sum_j\frac{\partial f}{\partial\xi_j}d\xi_j$ for $f\in\mathcal{O}_m\langle n\rangle$, and $d^2=0$. As usual, for any $X\in W(m|n)$ one defines a derivation ι_X (contraction along X) of the superalgebra $\Omega(m|n)$ by the properties (here x stands for x and ξ): $p(\iota_X)=p(X)+\bar{1}$, $\iota_X(x_j)=0$, $\iota_X(dx_j)=(-1)^{p(X)}X(x_j)$. The action of any $X\in W(m|n)$ on $\mathcal{O}_m\langle n\rangle$ extends in a unique way to the action by a derivation of $\Omega(m|n)$ such that $[X,d]=0$. This is called Lie's derivative and is usually denoted by L_X, but we shall write X in place of L_X unless confusion may arise. One has the usual Cartan's formula for this action: $L_X=[d,\iota_X]$.

Using this action, one can define super-analogs of the Hamiltonian and contact Lie algebras for any $n\in\mathbb{Z}_+$:

$$H(m|n) \;=\; \{X\in W(m|n)|X\omega_s=0\}\,,$$

where $m=2k$ and $\omega_s=\sum_{i=1}^k dx_i\wedge dx_{k+i}+\sum_{j=1}^n(d\xi_j)^2$,

$$K(m|n) \;=\; \{X\in W(m|n)|X\omega_c=f\omega_c\}\,,$$

where $m=2k+1$, $\omega_c=dx_m+\sum_{i=1}^k x_idx_{k+i}+\sum_{j=1}^n\xi_j\,d\xi_j$, and $f\in\mathcal{O}_m\langle n\rangle$.

Note that $W(0|n)$, $S(0|n)$ and $H(0|n)$ are finite-dimensional Lie superalgebras. The Lie superalgebras $W(0|n)$ and $S(0|n)$ are simple iff $n\geq 2$ and $n\geq 3$, respectively. However, $H(0|n)$ is not simple as its derived algebra $H'(0|n)$ has codimension 1 in $H(0|n)$, but $H'(0|n)$ is simple iff $n\geq 4$. Thus, in the Lie superalgebra

case the lists of simple finite- and infinite-dimensional algebras are much closer related than in the Lie algebra case.

These four series of Lie superalgebras are infinite-dimensional if $m \geq 1$, in which case they are simple except for $S(1|n)$. The derived algebra $S'(1|n)$ has codimension 1 in $S(1|n)$, and $S'(1|n)$ is simple iff $n \geq 2$.

Remarkably it turned out that the above four series do not exhaust all infinite-dimensional simple linearly compact Lie superalgebras (as has been suggested in [K2]). Far from it!

As was pointed out by several mathematicians, the Schouten bracket [SV] makes the space of polyvector fields on a m-dimensional manifold into a Lie superalgebra. The formal analog of this is the following fifth series of superalgebras, called by physicists the Batalin-Vilkoviski algebra (H stands here for "Hamiltonian" and O for "odd"):

$$HO(m|m) = \{X \in W(m|m) | X\omega_{os} = 0\},$$

where $\omega_{os} = \sum_{i=1}^{m} dx_i d\xi_i$ is an "odd" symplectic form. Furthermore, unlike in the $H(m|n)$ case, not all vector fields of $HO(m|n)$ have zero divergence, which gives rise to the sixth series:

$$SHO(m|m) = \{X \in HO(m|m) | \operatorname{div} X = 0\}.$$

The seventh series is the odd analog of $K(m|n)$ [ALS]:

$$KO(m|m+1) = \{X \in W(m|m+1) | X\omega_{oc} = f\omega_{oc}\},$$

where $\omega_{oc} = d\xi_{m+1} + \sum_{i=1}^{m} (\xi_i\, dx_i + x_i\, d\xi_i)$ is an odd contact form. One can take again the divergence 0 vector fields in $KO(m|m+1)$ in order to construct the eighth series, but the situation is more interesting. It turns out that for each $\beta \in \mathbb{C}$ one can define the deformed divergence $\operatorname{div}_\beta X$ [Ko], [K7], so that $\operatorname{div} = \operatorname{div}_0$ and

$$SKO(m|m+1;\beta) = \{X \in KO(m|m+1) | \operatorname{div}_\beta X = 0\}$$

is a subalgebra. The superalgebras $HO(m|m)$ and $KO(m|m+1)$ are simple iff $m \geq 2$ and $m \geq 1$, respectively. The derived algebra $SHO'(m|m)$ has codimension 1 in $SHO(m|m)$, and it is simple iff $m \geq 3$. The derived algebra $SKO'(m|m+1;\beta)$ is simple iff $m \geq 2$, and it coincides with $SKO(m|m+1;\beta)$ unless $\beta = 1$ or $\frac{m-2}{m}$ when it has codimension 1.

Some of the examples described above have simple "filtered deformations", all of which can be obtained by the following simple construction. Let L be a subalgebra of $W(m|n)$, where n is even. Then it happens in three cases that

$$L^{\sim} := (1 + \Pi_{j=1}^{n} \xi_j) L$$

is different from L, but is closed under bracket. As a result we get the following three series of superalgebras: $S^{\sim}(0|n)$ [K2], $SHO^{\sim}(m|m)$ [CK2] and $SKO^{\sim}(m|m+1;\frac{m+2}{m})$ [Ko] (the constructions in [Ko] and [CK2] were more complicated). We thus get the ninth and the tenth series of simple infinite-dimensional Lie superalgebras:

$$SHO^{\sim}(m|m), \quad m \geq 2, m \text{ even},$$
$$SKO^{\sim}(m|m+1;1+2/m), m \geq 3, m \text{ odd}.$$

A surprising discovery was made in [Sh1] where the existence of three exceptional simple infinite-dimensional Lie superalgebras was announced. The proof of the existence along with one more exceptional example was given in [Sh2]. An explicit construction of these four examples was given later in [CK3]. The fifth exceptional example was found in the work on conformal algebras [CK1] and independently in [Sh2]. (The alleged sixth exceptional example $E(2|2)$ of [K7] turned out to be isomorphic to $SKO(2|3;1)$ [CK3].)

Now I can state the first main theorem.

Theorem 1. [K7] *The complete list of simple infinite-dimensional linearly compact Lie superalgebras consists of ten series of examples described above and five exceptional examples:* $E(1|6)$, $E(3|6)$, $E(3|8)$, $E(4|4)$, *and* $E(5|10)$.

Here and before the notation $X(m|n)$ means that this superalgebra can be embedded in $W(m|n)$ and that this embedding is minimal possible; E stands for "exceptional".

Remark. The local classification of transitive primitive (i.e., leaving no invariant fibrations) actions on a (super)manifold M is equivalent to the classification of all "primitive" pairs (L, L_0), where L is a linearly compact Lie (super)algebra and L_0 is a maximal open subalgebra (such that $\dim L/L_0 = \dim M$) without non-zero ideals of L. If L is simple, choosing any maximal open subalgebra L_0, we get a primitive pair (L, L_0). One can show that if, in addition, L is a Lie algebra and $\dim L = \infty$, there exists a unique such L_0. According to Cartan's theorem, the remaining infinite-dimensional primitive pairs are the Lie algebras obtained from S_n and H_n by adding the Euler operator E. Using the structure results on general transitive linearly compact Lie algebras [G1], it is not difficult to reduce the classification of infinite-dimensional primitive pairs to the classification of simple infinite-dimensional linearly compact Lie algebras (cf. [G2]). Such a reduction is possible also in the Lie superalgebra case, but it is much more complicated for two reasons: (a) a simple linearly compact Lie superalgebra may have several maximal open subalgebras (see [CK3] for a classification), (b) construction of arbitrary primitive pairs in terms of simple primitive pairs is more complicated in the superalgebra case (see [K8]).

1.3. Here I will describe the classification of finite-dimensional simple Lie superalgebras. We already have four "non-classical" series: $W(0|n)$, $S(0|n)$, $H'(0|n)$ and $S^{\sim}(0|n)$. The four "classical" series are constructed as follows. Introduce the following even and odd Euler operators $E = \sum_i x_i \frac{\partial}{\partial x_i} + \sum_j \xi_j \frac{\partial}{\partial \xi_j} \in W(m|n)$ and $E_o = \sum_i x_i \frac{\partial}{\partial \xi_i} + \sum_i \xi_i \frac{\partial}{\partial x_i} \in W(m|m)$. Let

$$\begin{aligned} s\ell(m|n) &= \{X \in S(m|n) | [E, X] = 0\}, \\ spo(m|n) &= \{X \in H(m|n) | [E, X] = 0\}, \\ p(m|m) &= \{X \in SHO(m|m) | [E, X] = 0\}, \\ q(m|m) &= \{X \in W(m|m) | [E_o, X] = 0\}. \end{aligned}$$

The Lie algebras $s\ell_m = s\ell(m|0)$, $sp_m = spo(m|0)$, $so_n = spo(0|n)$ are simple. Furthermore, $s\ell(m|n)$ are simple for $m \neq n$, all $spo(m|n)$ and $p(m|m)$ $(m \geq 3)$ are simple. The superalgebra $s\ell(m|m)$ contains 1-dimensional ideal $\mathbb{C}E$ and $s\ell(m|m)/\mathbb{C}E$ is simple for $m \geq 2$. Finally, the derived algebra $q'(m|m)$ has codimension 1 in $q(m|m)$ and $q'(m|m)/\mathbb{C}E$ is simple for $m \geq 3$.

Theorem 2. [K2] *The complete list of simple finite-dimensional Lie superalgebras consists of eight series of examples described above, the exceptional Lie superalgebras* $F(4)$ *and* $G(3)$ *of dimension* 40 *and* 31, *respectively, a* 1*-parameter family of* 17*-dimensional exceptional Lie superalgebras* $D(2,1;a)$, *and the five exceptional Lie algebras.*

1.4. Plan of the proof of Theorem 1.

Step 1. Introduce Weisfeiler's filtration [W] of a simple linearly compact Lie superalgebra L. For that choose a maximal open subalgebra L_0 of L and a minimal subspace L_{-1} satisfying the properties: $L_{-1} \supsetneq L_0$, $[L_0, L_{-1}] \subset L_{-1}$. (Geometrically this corresponds to a choice of a primitive action of L and an invariant irreducible differential system.) The pair L_{-1}, L_0 can be included in a unique filtration: $L = L_{-d} \supsetneq L_{-d+1} \supset \cdots \supset L_{-1} \supset L_0 \supset L_1 \supset \cdots$, called *Weisfeiler's filtration* of *depth* d. (In the Lie algebra case, $d > 1$ only for K_m, when $d = 2$, but in the Lie superalgebra case, $d > 1$ in the majority of cases.) The associated to Weisfeiler's filtration $\mathbb{Z}$-graded Lie superalgebra is of the form $GrL = \Pi_{j \geq -d} \mathfrak{g}_j$, and has the following properties:

(G0) $\dim \mathfrak{g}_j < \infty$ (since $\operatorname{codim} L_0 < \infty$),
(G1) $\mathfrak{g}_{-j} = \mathfrak{g}_{-1}^j$ for $j \geq 1$ (by maximality of L_0),
(G2) $[x, \mathfrak{g}_{-1}] = 0$ for $x \in \mathfrak{g}_j$, $j \geq 0 \Rightarrow x = 0$ (by simplicity of L),
(G3) $\mathfrak{g}_0$-module $\mathfrak{g}_{-1}$ is irreducible (by choice of L_{-1}), and faithful (by (G2)).

Weisfeiler's idea was that property (G3) is so restrictive, that it should lead to a complete classification of $\mathbb{Z}$-graded Lie algebras satisfying (G0)–(G3). (Incidentally, the infinite-dimensionality of L and hence of GrL, since L is simple, is needed only in order to conclude that $\mathfrak{g}_1 \neq 0$.) This indeed turned out to be the case [K1]. In fact, my idea was to replace the condition of finiteness of the depth by finiteness of the growth, which allowed one to add to the Lie-Cartan list some new Lie algebras, called nowadays affine Kac-Moody algebras.

However, unlike in the Lie algebra case, it is impossible to classify all finite-dimensional irreducible faithful representations of Lie superalgebras. One needed a new idea to make this approach work.

Step 2. The main new idea is to choose L_0 to be invariant with respect to all inner automorphisms of L (meaning to contain all even ad-exponentiable elements of L). A non-trivial point is the existence of such L_0. This is proved by making use of the characteristic supervariety, which involves rather difficult arguments of Guillemin [G2].

Next, using a normalizer trick of Guillemin [G2], I prove, for this choice of L_0, the following very powerful restriction on the $\mathfrak{g}_0$-module $\mathfrak{g}_{-1}$ (at this point $\dim L = \infty$ is used):

(G4) $[\mathfrak{g}_0, x] = \mathfrak{g}_{-1}$ for any non-zero even element x of $\mathfrak{g}_{-1}$.

Step 3. Consider a faithful irreducible representation of a Lie superalgebra $\mathfrak{p}$ in a finite-dimensional vector space V. This representation is called *strongly transitive* if $\mathfrak{p} \cdot x = V$ for any non-zero even element $x \in V$. By properties (G0), (G3) and (G4), the $\mathfrak{g}_0$-module $\mathfrak{g}_{-1}$ is strongly transitive.

In order to demonstrate the power of this restriction, consider first the case when $\mathfrak{p}$ is a Lie algebra and V is purely even. Then the strong transitivity simply means that $V\backslash\{0\}$ is a single orbit of the Lie group P corresponding to $\mathfrak{p}$. It is rather easy to see that the only strongly transitive subalgebras $\mathfrak{p}$ of $g\ell_V$ are $g\ell_V$, $s\ell_V$, sp_V and csp_V. These four cases lead to GrL, where $L = W_n$, S_n, H_n and K_n, respectively.

In the super case the situation is much more complicated. First we consider the case of "inconsistent gradation", meaning that $\mathfrak{g}_{-1}$ contains a non-zero even element. The classification of such strongly transitive modules is rather long and the answer consists of a dozen series and a half dozen exceptions (see [K7], Theorem 3.1). Using similar restrictions on $\mathfrak{g}_{-2}, \mathfrak{g}_{-3}, \ldots$, we obtain a complete list of possibilities for $GrL_{\leq} := \oplus_{j \leq 0} \mathfrak{g}_j$, in the case when $\mathfrak{g}_{-1}$ contains non-zero even elements. It turns out that all but one exception are not exceptions at all, but correspond to the beginning members of some series. As a result, only $E(4|4)$ "survives".

Step 4. Next, we turn to the case of a consistent gradation, i.e., when $\mathfrak{g}_{-1}$ is purely odd. But then $\mathfrak{g}_0$ is an "honest" Lie algebra, having a faithful irreducible representation in $\mathfrak{g}_{-1}$ (condition (G4) becomes vacuous). An explicit description of such representations is given by the classical Cartan-Jacobson theorem. In this case I use the "growth" method developed in [K1] and [K2] to determine a complete list of possibilities for $GrL_{\leq}$. This case produces mainly the (remaining four) exceptions.

Step 5 is rather long and tedious [CK3] . For each $GrL_{\leq}$ obtained in Steps 3 and 4 we determine all possible "prolongations", i.e., infinite-dimensional $\mathbb{Z}$-graded Lie superalgebras satisfying (G2), whose negative part is the given $GrL_{\leq}$.

Step 6. It remains to reconstruct L from GrL, i.e., to find all possible filtered simple linearly compact Lie superalgebras L with given GrL (such an L is called a simple filtered deformation of GrL). Of course, there is a trivial filtered deformation: $GrL := \Pi_{j \geq -d} \mathfrak{g}_j$, which is simple iff GrL is. It is proved in [CK2] by a long and tedious calculation that only $SHO(m|m)$ for m even ≥ 2 and $SKO(m|m+1; \frac{m+2}{m})$ for m odd ≥ 3 have a non-trivial simple filtered deformation, which are the ninth and tenth series. It would be nice to have a more conceptual proof. Recall that $SH0(m|m)$ is not simple, though it does have a simple filtered deformation. Note also that in the Lie algebra case all filtered deformations are trivial.

1.5. Plan of the proof of Theorem 2.

The key idea is the same as in the proof of Theorem 1. Choose a maximal subalgebra L_0 of a simple finite-dimensional Lie superalgebra L containing the even part of L. It is easy to see then that $L_{-1} = L$, so that the corresponding Weisfeiler's filtration has depth 1. Hence GrL has the form: $GrL = \oplus_{j=-1}^{N} \mathfrak{g}_j$. This gradation is consistent and, of course, satisfies conditions (G0)–(G3). There are two cases.

Case 1. $N \geq 1$. Then we apply the growth method (as in Step 4 of Sec. 1.4.) to obtain a complete list of possibilities for $GrL_{\leq}$. Then, as in Step 5 of Sec. 1.4. we determine all prolongations of each $GrL_{\leq}$ (all of them will be subalgebras of $W(0, \dim \mathfrak{g}_1)$), and all filtered deformations of these prolongations. This case produces all "non-classical" series, and also $s\ell(m|n)$, $spo(m|2)$ and $p(m|m)$.

Case 2. $N = 0$. Then $L_{\bar{0}}$ is a semisimple Lie algebra and its representation in $L_{\bar{1}}$ is irreducible. The Killing form on L is either non-degenerate, in which case we apply the standard Killing-Cartan techniques, or it is identically zero. In the latter case one uses Dynkin's index to find all possibilities for the $L_{\bar{0}}$-module $L_{\bar{1}}$.

1.6. In order to describe the construction of the exceptional infinite-dimensional Lie superalgebras (given in [CK3]), I need to make some remarks. Let $\Omega_m = \Omega(m|0)$ be the algebra of differential forms over $\mathcal{O}_m$, let Ω_m^k denote the space of forms of degree k, and $\Omega_{m,c\ell}^k$ the subspace of closed forms. For any $\lambda \in \mathbb{C}$ the representation of W_m on Ω_m^k can be "twisted" by letting

$$X \mapsto L_X + \lambda \operatorname{div} X, \quad X \in W_m\,,$$

to get a new W_m-module, denoted by $\Omega_m^k(\lambda)$ (the same can be done for $W(m|n)$). Obviously, $\Omega_m^k(\lambda) = \Omega_m^k$ when restricted to S_m. Then we have the following obvious W_m-module isomorphisms: $\Omega_m^0 \simeq \Omega_m^m(-1)$ and $\Omega_m^0(1) \simeq \Omega_m^m$. Furthermore, the map $X \mapsto \iota_X(dx_1 \wedge \ldots \wedge dx_m)$ gives the following W_m-module and S_m-module isomorphisms:

$$W_m \simeq \Omega_m^{m-1}(-1), \quad S_m \simeq \Omega_{m,c\ell}^{m-1}\,.$$

We shall identify the representation spaces via these isomorphisms.

The simplest is the construction of the largest exceptional Lie superalgebra $E(5|10)$. Its even part is the Lie algebra S_5, its odd part is the space of closed 2-forms $\Omega_{5,c\ell}^2$. The remaining commutators are defined as follows for $X \in S_5$, $\omega, \omega' \in \Omega_{5,c\ell}^2$:

$$[X, \omega] = L_X\omega, \quad [\omega, \omega'] = \omega \wedge \omega' \in \Omega_{5,c\ell}^4 = S_5\,.$$

Each quintuple of integers $(a_1, a_2, \ldots, a_5)$ such that $a = \sum_i a_i$ is even, defines a $\mathbb{Z}$-gradation of $E(5|10)$ by letting:

$$\deg x_i = -\frac{\partial}{\partial x_i} = a_i, \quad \deg dx_i = a_i - \tfrac{1}{4}a\,.$$

The quintuple $(2, 2, \ldots, 2)$ defines the (only) consistent $\mathbb{Z}$-gradation, which has depth 2: $E(5|10) = \Pi_{j \geq -2}\mathfrak{g}_j$, and one has:

$$\mathfrak{g}_0 \simeq s\ell_5 \text{ and } \mathfrak{g}_{-1} \simeq \Lambda^2\mathbb{C}^5, \quad \mathfrak{g}_{-2} \simeq \mathbb{C}^{5*} \text{ as } \mathfrak{g}_0\text{-modules.}$$

Furthermore, $\Pi_{j \geq 0}\mathfrak{g}_j$ is a maximal open subalgebra of $E(5|10)$ (the only one which is invariant with respect to all automorphisms). There are three other maximal open subalgebras in $E(5|10)$, associated to $\mathbb{Z}$-gradations corresponding to quintuples $(1, 1, 1, 1, 2)$, $(2, 2, 2, 1, 1)$ and $(3, 3, 2, 2, 2)$, and one can show that these four are all, up to conjugacy, maximal open subalgebras (cf. [CK3]).

Another important $\mathbb{Z}$-gradation of $E(5|10)$, which is, unlike the previous four, by infinite-dimensional subspaces, corresponds to the quintuple $(0,0,0,1,1)$ and has depth 1: $E(5|10) = \Pi_{\lambda \geq -1} \mathfrak{g}^\lambda$. One has: $\mathfrak{g}^0 \simeq E(3|6)$ and the $\mathfrak{g}^\lambda$ form an important family of irreducible $E(3|6)$-modules [KR2]. The consistent $\mathbb{Z}$-gradation of $E(5|10)$ induces that of $\mathfrak{g}^0 : E(3|6) = \Pi_{j \geq -2} \mathfrak{a}_j$, where

$$\mathfrak{a}_0 \simeq s\ell_3 \oplus s\ell_2 \oplus g\ell_1, \quad \mathfrak{a}_{-1} \simeq \mathbb{C}^3 \boxtimes \mathbb{C}^2 \boxtimes \mathbb{C}, \quad \mathfrak{a}_{-2} \simeq \mathbb{C}^3 \boxtimes \mathbb{C} \boxtimes \mathbb{C}.$$

A more explicit construction of $E(3|6)$ is as follows [CK3]: the even part is $W_3 + \Omega_3^0 \otimes s\ell_2$, the odd part is $\Omega_3^1(-\frac{1}{2}) \otimes \mathbb{C}^2$ with the obvious action of the even part, and the bracket of two odd elements is defined as follows:

$$[\omega \otimes u, \omega' \otimes v] = (\omega \wedge \omega') \otimes (u \wedge v) + (d\omega \wedge \omega' + \omega \wedge d\omega') \otimes (u \cdot v)\,.$$

Here the identifications $\Omega_3^2(-1) = W_3$ and $\Omega_3^0 = \Omega_3^3(-1)$ are used.

The gradation of $E(5|10)$ corresponding to the quintuple $(0,1,1,1,1)$ has depth 1 and its 0^{th} component is isomorphic to $E(1|6)$ (cf. [CK3]).

The construction of $E(4|4)$ is also very simple [CK3]: The even part is W_4, the odd part is $\Omega_4^1(-\frac{1}{2})$ and the bracket of two odd elements is:

$$[\omega, \omega'] = d\omega \wedge \omega' + \omega \wedge d\omega' \in \Omega_4^3(-1) = W_4\,.$$

The construction of $E(3|8)$ is slightly more complicated, and we refer to [CK3] for details.

1.7. All exceptional simple finite-dimensional Lie superalgebras (including the exceptional Lie algebras) are obtained as special cases of the following important construction [K2]. Let $I = \{1, \dots, r\}$ and let $I_{\overline{1}}$ be a subset of I, $I_{\overline{0}} = I \backslash I_{\overline{1}}$. Let $A = (a_{ij})_{i,j \in I}$ be a matrix over $\mathbb{C}$. We associate to the pair $(A, I_{\overline{1}})$ a Lie superalgebra $\mathfrak{g}(A, I_{\overline{1}})$ as follows. Let $\tilde{\mathfrak{g}}(A, I_{\overline{1}})$ be the Lie superalgebra on generators e_i, f_i, h_i $(i \in I)$ of parity $p(h_i) = \overline{0}$ for $i \in I$, $p(e_i) = p(f_i) = \alpha \in \{\overline{0}, \overline{1}\}$ for $i \in I_\alpha$, and the following standard relations:

$$[h_i, h_j] = 0, [e_i, f_j] = \delta_{ij} h_i\,,\ [h_i, e_j] = a_{ij} e_j, [h_i, f_j] = -a_{ij} f_j.$$

Define a $\mathbb{Z}$-gradation $\tilde{\mathfrak{g}}(A, I_{\overline{1}}) = \oplus_{j \in \mathbb{Z}} \tilde{\mathfrak{g}}_j$ by letting $\deg h_i = 0$, $\deg e_i = -\deg f_i = 1$. Then $\tilde{\mathfrak{g}}_0$ is the $\mathbb{C}$-span of $\{h_i\}_{i \in I}$, and we denote by $J(A, I_{\overline{1}})$ the sum of all $\mathbb{Z}$-graded ideals of $\tilde{\mathfrak{g}}(A, I_{\overline{1}})$ that intersect $\tilde{\mathfrak{g}}_0$ trivially. We let

$$\mathfrak{g}(A, I_{\overline{1}}) = \tilde{\mathfrak{g}}(A, I_{\overline{1}}) / J(A, I_{\overline{1}})\,.$$

Of course, if A is the Cartan matrix of a simple finite-dimensional Lie algebra $\mathfrak{g}$, then $\mathfrak{g} \simeq \mathfrak{g}(A, \emptyset)$, the ideal $J(A, \emptyset)$ being generated by "Serre relations". Likewise, generalized Cartan matrices give rise to Kac-Moody Lie algebras [K3].

Consider the following matrices $(a \in \mathbb{C} \backslash \{0, -1\})$:

$$D_a = \begin{bmatrix} 0 & -1 & -a \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{bmatrix}, F = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 2 & -2 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}, G = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 2 & -3 \\ 0 & -1 & 2 \end{bmatrix}.$$

Then $D(2,1;\alpha) \simeq \mathfrak{g}(D_\alpha, \{1\})$, $F(4) \simeq \mathfrak{g}(F, \{1\})$ and $G(3) \simeq \mathfrak{g}(G, \{1\})$. Note, however, that unlike in the Lie algebra case, "inequivalent" pairs $(A, I_{\overline{1}})$ may produce isomorphic Lie superalgebras. For example, in the cases $D(2,1;a)$, $F(4)$ and $G(3)$ there are 2, 6 and 4 such pairs, respectively.

Finite-dimensional simple Lie superalgebras that are isomorphic to $\mathfrak{g}(A, I_{\overline{1}})$ for some matrix A are called *basic* (they will play an important role in the next parts of the talk). The remaining basic simple Lie superalgebras (that are not Lie algebras) are $s\ell(m|n)/\delta_{m,n}\mathbb{C}E$ and $spo(m|n)$.

2. A classification of superconformal algebras

Superconformal algebras have been playing an important role in superstring theory and in conformal field theory. Here I will explain how to apply Theorem 1 to the classification of "linear" superconformal algebras. By a ("linear") superconformal algebra I mean a Lie superalgebra $\mathfrak{g}$ spanned by coefficients of a finite collection F of fields such that the following two properties hold:

(1) for $a, b \in F$ the singular part of OPE is finite, i.e.,

$$[a(z), b(w)] = \sum_j c_j(w)\partial_w^j \delta(z-w) \quad \text{(a finite sum), where all } c_j(w) \in \mathbb{C}[\partial_w]F,$$

(2) $\mathfrak{g}$ contains no non-trivial ideals spanned by coefficients of fields from a $\mathbb{C}[\partial_w]$-submodule of $\mathbb{C}[\partial_w]F$.

(Recall that a field is a formal expression $a(z) = \sum_{n\in\mathbb{Z}} a_n z^n$, where $a_n \in \mathfrak{g}$ and z is an indeterminate, and $\delta(z-w) = z^{-1}\sum_{n\in\mathbb{Z}}(w/z)^n$ is the formal δ-function. See [K5] for details.)

This problem goes back to the physics paper [RS], some progress in its solution was made in [K6] and a complete solution was given in [FK]. (A complete classification even in the "quadratic" case seems to be a much harder problem, see [FL] and Section 4 below for some very interesting examples.) The simplest example is the loop algebra $\tilde{\mathfrak{g}} = \mathbb{C}[x, x^{-1}] \otimes \mathfrak{g}$ (= centerless affine Kac-Moody superalgebra), where $\mathfrak{g}$ is a simple finite-dimensional Lie superalgebra. Then $F = \{a(z) = \sum_{n\in\mathbb{Z}}(x^n \otimes a)z^{-n-1}\}_{a\in\mathfrak{g}}$, and $[a(z), b(w)] = [a,b](w)\delta(z-w)$. The next example is the Lie algebra $\mathrm{Vect}\,\mathbb{C}^\times$ of regular vector fields on $\mathbb{C}^\times$ (= centerless Virasoro algebra); F consists of one field, the Virasoro field $L(z) = -\sum_{n\in\mathbb{Z}}(x^n \frac{d}{dx})z^{-n-1}$, and $[L(z), L(w)] = \partial_w L(w)\delta(z-w) + 2L(w)\delta'_w(z-w)$.

One of the main theorems of [DK] states that these are all examples in the Lie algebra case. The strategy of the proof is the following. Let $\partial = \partial_z$ and consider the (finitely generated) $\mathbb{C}[\partial]$-module $R = \mathbb{C}[\partial]F$. Define the "$\lambda$-bracket" $R \otimes R \to \mathbb{C}[\lambda] \otimes R$ by the formula: $[a_\lambda b] = \sum_j \lambda^j c_j$. This satisfies the axioms of a *conformal (super)algebra* (see [DK], [K5]), similar to the Lie (super)algebra axioms:

(i) $[\partial a_\lambda b] = -\lambda[a_\lambda b]$, $[a_\lambda \partial b] = (\partial + \lambda)[a_\lambda b]$,
(ii) $[a_\lambda b] = -(-1)^{p(a)p(b)}[b_{-\lambda-\partial}a]$,
(iii) $[a_\lambda [b_\mu c]] = [[a_\lambda b]_{\lambda+\mu} c] + (-1)^{p(a)p(b)}[b_\mu [a_\lambda c]]$.

The main observation of [DK] is that a conformal (super)algebra is completely determined by the Lie (super)algebra spanned by all coefficients of negative powers of z of the fields $a(z)$ from F, called the *annihilation algebra*, along with an even surjective derivation of the annihilation algebra. Furthermore, apart from the case of current algebras, the completed annihilation algebra turns out to be an infinite-dimensional simple linearly compact Lie (super)algebra of growth 1. Since in the Lie algebra case the only such example is W_1, the proof is finished.

In the superalgebra case the situation is much more interesting since there are many infinite-dimensional simple linearly compact Lie superalgebras of growth 1. By Theorem 1, the complete list is as follows:

$$W(1|N), \quad S'(1|N), \quad K(1|N) \text{ and } E(1|6)\,.$$

In all cases, except the second, there is a unique, up to conjugacy, even surjective derivation, hence a unique corresponding superconformal algebra. They are denoted by $W_{(N)}$, $K_{(N)}$ if $N \neq 4$, $K'_{(4)}$ and $CK_{(6)}$, respectively. The Lie superalgebras $W_{(N)}$ and $K_{(N)}$ are constructed in the same way as $W(1|N)$ and $K(1|N)$, except that one replaces $\mathcal{O}_1\langle N\rangle$ by $\mathbb{C}[x,x^{-1}]\langle N\rangle$. The construction of the exceptional superconformal algebra $CK_{(6)}$ is more difficult, and may be found in [CK1] or [K6]. However, $S'(1|N)$ has two families of even surjective derivations. The corresponding superconformal algebras are derived algebras of

$$S_{(N),\epsilon,a} = \{X \in W_{(N)} | \operatorname{div}(e^{ax}(1+\epsilon\xi_1 \dots \xi_N)X) = 0\}\,, \quad a \in \mathbb{C},\ \epsilon = \pm 1\,.$$

Thus, one obtains the following theorem.

Theorem 3. [FK] *A complete list of superconformal algebras consists of loop algebras $\tilde{\mathfrak{g}}$, where $\mathfrak{g}$ is a simple finite-dimensional Lie superalgebra, and of Lie superalgebras ($N \in \mathbb{Z}_+$): $W_{(N)}$, $S'_{(N+2),\epsilon,a}$ (N even and $a = 0$ if $\epsilon = 1$), $K_{(N)} (N \neq 4)$, $K'_{(4)}$, and $CK_{(6)}$.*

Note that the first members of the above series are well-known superalgebras: $W_{(0)} \simeq K_{(0)}$ is the Virasoro algebra, $K_{(1)}$ is the Neveu-Schwarz algebra, $K_{(2)} \simeq W_{(1)}$ is the $N = 2$ algebra, $K_{(3)}$ is the $N = 3$ algebra, $S'_{(2)}{=}S'_{(2),0,0}$ is the $N = 4$ algebra, $K'_{(4)}$ is the big $N = 4$ algebra (all centerless). These algebras, along with $W_{(2)}$ and $CK_{(6)}$ are the only superconformal algebras for which all fields are primary with positive conformal weights [K6]. It is interesting to note that all of them are contained in $CK_{(6)}$, which consists of 32 fields, the even ones are the Virasoro fields and 15 currents that form $\tilde{so}_6$, and the odd ones are 6 and 10 fields of conformal weight 3/2 and 1/2, respectively. Here is the table of (some) inclusions, where in square brackets the number of fields is indicated:

$$\begin{array}{ccccccc} CK_{(6)}[32] & & \supset & W_{(2)}[12] & \supset W_{(1)} = K_{(2)}[4] \supset K_{(1)}[2] \supset \mathrm{Vir}[1] \\ \cup & & & \cup & & & \\ K_{(3)}[8] \subset K'_{(4)}[16] & & & S'_{(2),\epsilon,a}[8] & & & \end{array}\,.$$

All of these Lie superalgebras have a unique non-trivial central extension, except for K'_4 that has three [KL] and $CK_{(6)}$ that has none. All other Lie superalgebras

listed by Theorem 2 have no non-trivial central extensions. (The presence of a central term is necessary for the existence of interesting representations and the construction of an interesting conformal field theory.)

3. Representations of affine superalgebras and "almost" modular forms

3.1. Finite-dimensional irreducible representations of simple finite-dimensional Lie superalgebras are much less understood than in the Lie algebra case, the main reason being the occurence of isotropic roots in the super case. (A review may be found in the proceedings of the last ICM, see [Se].) The natural analogues of these representations in the case of affine (super)algebras are the integrable highest weight modules.

Let us first recall the basic definitions in the Lie algebra case, i.e., for an affine Kac-Moody algebra $\hat{\mathfrak{g}}$ [K3]. Let $\mathfrak{g}$ be a finite-dimensional simple Lie algebra and let $(.\,|\,.)$ be an invariant symmetric bilinear form on $\mathfrak{g}$ normalized by the condition that $(\alpha|\alpha) = 2$ for a long root α ($(a|b) = \mathrm{tr}ab$ in the case $\mathfrak{g} = s\ell_m$). Recall that the associated *affine algebra* is

$$\hat{\mathfrak{g}} = (\mathbb{C}[x, x^{-1}] \otimes_{\mathbb{C}} \mathfrak{g}) \oplus \mathbb{C}K \oplus \mathbb{C}D$$

with the following commutation relations ($a, b \in \mathfrak{g}$; $m, n \in \mathbb{Z}$ and $a(m)$ stands for $x^m \otimes a$):

$$[a(m), b(n)] = [a, b]\,(m+n) + m\delta_{m,-n}(a|b)K\,,\ [D, a(m)] = ma(m), [K, \hat{\mathfrak{g}}] = 0\,.$$

Note that the derived algebra $\hat{\mathfrak{g}}'$ is a central extension (by $\mathbb{C}K$) of the loop algebra $\tilde{\mathfrak{g}}$ that has made an appearance in Section 2. (It is also isomorphic to $\mathfrak{g}(\hat{A})$, where $\hat{A}$ is the extended Cartan matrix of $\mathfrak{g}$, cf. Sec. 1.7..) As we shall see, without central extension one loses all interesting representations. In any irreducible $\hat{\mathfrak{g}}$-modle V one has: $K = kI_V$; the number k is called the *level* of V. The scaling element D is necessary for the convergence of characters.

Choose a Cartan subalgebra $\mathfrak{h}$ of $\mathfrak{g}$ and let $\mathfrak{g} = \mathfrak{h} \oplus (\oplus_{\alpha\in\Delta}\mathfrak{g}_\alpha)$ be the root space decomposition, where $\mathfrak{g}_\alpha$ denotes the root space attached to a root $\alpha \in \Delta \subset \mathfrak{h}^*$. Let $\hat{\mathfrak{h}} = \mathfrak{h} + \mathbb{C}K + \mathbb{C}D$ be the Cartan subalgebra of $\hat{\mathfrak{g}}$, and, as before, let $\mathfrak{g}_\alpha(m) = x^m \otimes \mathfrak{g}_\alpha$. We extend the invariant bilinear form from $\mathfrak{h}$ to a symmetric bilinear form on $\hat{\mathfrak{h}}$ by letting $(\mathfrak{h}|\mathbb{C}K + \mathbb{C}D) = 0$, $(K|K) = (D|D) = 0$, $(K|D) = 1$, and identify $\hat{\mathfrak{h}}$ with $\hat{\mathfrak{h}}^*$ via this form.

A $\hat{\mathfrak{g}}$-module V is called *integrable* if the following two properties hold:

(M1) $\hat{\mathfrak{h}}$ is diagonizable on V,
(M2) for each $\alpha \in \Delta$ and $m \in \mathbb{Z}$, $\mathfrak{g}_\alpha(m)$ is locally finite on V.

Choose a set of positive roots $\Delta^+ \subset \Delta$, and let $\mathfrak{n}^+ = \oplus_{\alpha\in\Delta^+}\mathfrak{g}_\alpha$, $\hat{\mathfrak{n}}^+ = \mathfrak{n}^+ + \sum_{n\geq 1} x^n \otimes \mathfrak{g}$. For each $\Lambda \in \hat{\mathfrak{h}}^*$ one defines an *irreducible highest weight module*

$L(\Lambda)$ over $\hat{\mathfrak{g}}$ as the (unique) irreducible $\hat{\mathfrak{g}}$-module for which there exists a non-zero vector v_Λ such that

$$hv_\Lambda = \Lambda(h)v_\Lambda \text{ for all } h \in \hat{\mathfrak{h}}\,,\ \hat{\mathfrak{n}}_+ v_\Lambda = 0\,.$$

Without loss of generality we shall let $\Lambda(D) = 0$; then the spectrum of $-D$ on $L(\Lambda)$ is $\mathbb{Z}_+$.

Integrable highest weight modules over affine Lie algebras (they are automatically irreducible) attracted a lot of attention in the past few decades both of mathematicians and of physicists (some aspects of the theory are discussed in [K3], [Wa2].) Here I will only mention some relevant to the talk facts. First, the level k of such a module is a non-negative integer (and $k = 0$ iff $\dim L(\Lambda) = 1$), and there is a finite number of them for each k. One of the most remarkable properties of these modules is modular invariance, which I explain below.

3.2. Let us coordinatize $\hat{\mathfrak{h}}$ by letting

$$(\tau, z, t) = 2\pi i(z - \tau D + tK)\,,$$

where $z \in \mathfrak{h}$, $\tau, t \in \mathbb{C}$, and define the *character* of the $\hat{\mathfrak{g}}$-module $L(\Lambda)$ by:

$$ch_\Lambda(\tau, z, t) = \mathrm{tr}_{L(\Lambda)} e^{2\pi i(z - \tau D + tK)}\,.$$

If $L(\Lambda)$ is integrable, then $ch_\Lambda(\tau, z, t)$ is a holomorphic function for $(\tau, z, t) \in \mathcal{H} \times \mathfrak{h} \times \mathbb{C}$, where $\mathcal{H} = \{\tau \in \mathbb{C} \,|\, \mathrm{Im}\ \tau > 0\}$.

Recall the following well-known action of $SL_2(\mathbb{Z})$ on $\mathcal{H} \times \mathfrak{h} \times \mathbb{C}$:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot (\tau, z, t) = \left(\frac{a\tau + b}{c\tau + d}\,,\ \frac{z}{c\tau + d}\,,\ t - \frac{c(z|z)}{2(c\tau + d)} \right)\,.$$

Then it turns out that there exists an explicit rational number m_Λ, called *modular anomaly* (see [K3], (12.7.5)) such that the normalized character $\chi_\Lambda = e^{2\pi i m_\Lambda \tau} ch_\Lambda$ of an integrable $L(\Lambda)$ is invariant with respect to a congruence subgroup of $SL_2(\mathbb{Z})$ (see [K3], Chapter 13). If ch_Λ has this property, one says that $L(\Lambda)$ is *modular invariant.*

3.3. It turns out that $L(\Lambda)$ is modular invariant for a much wider (than integrable) collection of Λ's, called *admissible*, defined by the condition [KW1], [KW2], [K4]:

$$2(\Lambda + \hat{\rho}|\alpha)/(\alpha|\alpha) \in \mathbb{Q}\backslash\{0, -1, -2, \ldots\} \text{ for all } \alpha \in \hat{\Delta}^+ \text{ such that } (\alpha|\alpha) \neq 0\,.$$

(The conjecture of Wakimoto and myself is that these are all modular invariant $L(\Lambda)$.) Here $\hat{\Delta}^+$ is the set of positive roots of $\hat{\mathfrak{g}}$ corresponding to $\hat{\mathfrak{n}}^+$: $\hat{\Delta}^+ = \Delta^+ \cup \{\alpha + nK | \alpha \in \Delta, n \geq 1\} \cup \{nK | n \geq 1\}$, and $\hat{\rho} \in \hat{\mathfrak{h}}^*$ is a vector satisfying $(\hat{\rho}|\alpha_i) = \frac{1}{2}(\alpha_i|\alpha_i)$ for $i = 0, 1, \ldots, r$, where $\Pi = \{\alpha_1, \ldots, \alpha_r\}$ are simple roots of Δ^+, $\hat{\Pi} = \{\alpha_0 := K - \theta\} \cup \Pi$ are simple roots of $\hat{\Delta}^+$, θ is the highest root of Δ^+.

I shall describe explicitly the most important class of them, called *principal admissible.* Fix a positive integer u and let $\hat{\Pi}_u = \{uK - \theta\} \cup \Pi$. Let $k = v/u$ be

a rational number, such that $v \in \mathbb{Z}$ is realtively prime to u and $u(k+h^\vee) \geq h^\vee$. Here $h^\vee$ is the dual Coxeter number (defined in a more general Lie superalgebra context further on). Let W be the Weyl group of $\mathfrak{g}$ and let $P^\vee = \{\lambda \in \mathfrak{h} | (\lambda|\alpha) \in \mathbb{Z}$ for all $\alpha \in \Delta\}$. For each $\alpha \in P^\vee$ define a translation $t_\alpha \in \text{End } \hat{\mathfrak{h}}$ by the formula $t_\alpha(\lambda) = \lambda + (\lambda|K)\alpha - ((\lambda|\alpha) + \frac{(\lambda|K)}{2}(\alpha|\alpha))K$. Pick an element $\hat{w} = t_\beta w$, where $w \in W$, such that $\hat{w}\hat{\Pi}_u \subset \hat{\Delta}^+$ (these are all subsets of $\hat{\Delta}^+$ isomorphic to $\hat{\Pi}$). Let Λ^0 be an integrable highest weight of level $u(k+h^\vee) - h^\vee$. Then

$$\Lambda = \hat{w}(\Lambda^0 + \hat{\rho} - (u-1)(k+h^\vee)D) - \hat{\rho}$$

is a *principal admissible weight* of level k. The character of the corresponding $L(\Lambda)$ is given by the following formula [KW1]–[KW2]:

$$\left(\hat{R}ch_{L(\Lambda)}\right)(\tau,z,t) = \left(\hat{R}ch_{L(\Lambda^0)}\right)\left(u\tau, w^{-1}(z+\tau\beta)\,,\, u^{-1}\left(t + (z|\beta) + \tfrac{1}{2}\tau(\beta|\beta)\right)\right)\,, \tag{Ch}$$

where $\hat{R} = e^{\hat{\rho}}\prod_{\alpha\in\hat{\Delta}^+}(1-e^{-\alpha})^{\text{mult}\,\alpha}$ is the Weyl denominator function for $\hat{\mathfrak{g}}$, and $\text{mult}\,\alpha = 1$ except for $\text{mult}\,nK = r$ for all n (note that this formula is a tautology if $\Lambda = \Lambda^0$ is integrable; this happens iff $u = 1$). Recall that $ch_{L(\Lambda^0)}$ is given by the Weyl-Kac character formula [K3].

3.4. Let now $\mathfrak{g} = \mathfrak{g}(A)$ be a basic simple finite-dimensional Lie superalgebra (see Sec. 1.7). Then $\mathfrak{g}$ carries a unique, up to a constant factor, non-degenerate invariant bilinear form B ("invariant" means that $B([a,b],c) = B(a,[b,c])$). Let $\mathfrak{h} = \sum_{i=1}^r \mathbb{C}h_i$ be the Cartan subalgebra of $\mathfrak{g}$, $\mathfrak{n}^+$ the subalgebra of $\mathfrak{g}$ generated by all e_i, $\Delta^+ \subset \mathfrak{h}^*$ the set of positive roots (i.e., roots of $\mathfrak{h}$ in $\mathfrak{n}^+$), $\Delta = \Delta^+ \cup -\Delta^+$ the set of all roots, $\Delta_{\bar{0}}$ and $\Delta_{\bar{1}}$ the sets of even and odd roots, $\{\alpha_1,\dots,\alpha_r\} \subset \Delta^+$ the set of simple roots $(\alpha_i(h_j) = a_{ji})$, $\theta \in \Delta^+$ the highest root. Define $\rho \in \mathfrak{h}^*$ by $B(\rho,\alpha_i) = \frac{1}{2}B(\alpha_i,\alpha_i)$, $i = 1,\dots,r$. Then $h_B^\vee = B(\rho,\rho) + \frac{1}{2}B(\theta,\theta)$ is the eigenvalue of the Casimir operator in the adjoint representation.

If $\mathfrak{h}_B^\vee \neq 0$, we let $\Delta_0^\# = \{\alpha \in \Delta_{\bar{0}} \,|\, h_B^\vee B(\alpha,\alpha) > 0\}$. If $h_B^\vee = 0$, which happens for $\mathfrak{g} = s\ell(m|m)/\mathbb{C}E$, $spo(2m|2m+2)$ and $D(2,1;a)$, we take for $\Delta_0^\#$ the sets of roots of the subalgebra $s\ell_m$, so_{2m+2} and $s\ell_2 \oplus s\ell_2$, respectively. Let $W^\#$ denote the subgroup of the Weyl group of $\mathfrak{g}_{\bar{0}}$ generated by reflections with repsect to all $\alpha \in \Delta_0^\#$. Denote by $(.\,|\,.)$ the invariant bilinear form on $\mathfrak{g}$ normalized by the condition $(\alpha|\alpha) = 2$ for the longest root $\alpha \in \Delta_0^\#$. The corresponding to this form number $h^\vee = h^\vee_{(.\,|\,.)}$ is called *the dual Coxeter number.* (For example, this number equals $|m-n|$ for $s\ell(m|n)$, $\frac{1}{2}(m-n)+1$ for $spo(m|n)$ with $m \geq n-2$, 30 for E_8, 3 for $F(4)$, 2 for $G(3)$.)

3.5. Define the affine superalgebra $\hat{\mathfrak{g}}$ associated to the Lie superalgebra $\mathfrak{g}$ in exactly the same way as in the Lie algebra case. The highest weight $\hat{\mathfrak{g}}$-modules $L(\Lambda)$ are defined in the same way too. The *integrability* of a $\hat{\mathfrak{g}}$-module V is defined by (M1), (M2) with Δ replaced by $\Delta_0^\#$, and

(M3) V is locally $\mathfrak{g}$-finite. ((M1) and (M2) imply (M3), but there are no integrable $L(\Lambda)$ in the super case if one doesn't replace Δ by $\Delta_0^{\#}$.)

Integrable $\hat{\mathfrak{g}}$-modules $L(\Lambda)$ were classified in [KW4]. However, very little is known about their characters. The following example shows that modular invariance fails already in the simplest case $\mathfrak{g} = s\ell(2|1), \Lambda = D$. In this case

$$e^{-D}ch_{L(D)} = \prod_{n=1}^{\infty}((1-q^n)^{-1})(1+z_1q^n)(1+z_1^{-1}q^{n-1})A(z_1^{-1}, z_2q^{1/2}, q)\,, \qquad \text{(A)}$$

where $z_i = e^{\epsilon_i+\epsilon_3}$ ($i = 1, 2$; ϵ_i is the standard basis of the space of 3×3 diagonal matrices), $q = e^{2\pi i\tau}$, and

$$A(x, z, q) = \sum_{n\in\mathbb{Z}} \frac{q^{n^2/2}z^n}{1+xq^n}\,.$$

The function $A(x, z, q)$ converges to a meromorphic function in the domain $x, y, z \in \mathbb{C}$, $|q| < 1$, and is called Appell's function. Since the first factor in (A) has the modular invariance property and $A(x, z, q)$ doesn't have it, we see that $L(D)$ is not modular invariant. We call the Appell function an *almost modular form* since it is a section of a rank 2 vector bundle on an elliptic curve for each $\tau \in \mathcal{H}$ [P] (whereas modular forms are sections of rank 1 vector bundles on it).

We call a weight $\Lambda \in \hat{\mathfrak{h}}^*$ of the Lie superalgebra $\hat{\mathfrak{g}}$ *admissible* (resp. *principal admissible*) if Λ is admissible (resp. principle admissible) for the affine Lie algebra associated to a semisimple Lie algebra with root system $\Delta_0^{\#}$, and we conjecture [KW4] that formula (Ch) still holds, where $\hat{R}$ in the super case is defined by $\hat{R} = e^{\hat{\rho}}\prod_{\alpha\in\hat{\Delta}_+}(1-(-1)^{p(\alpha)}e^{-\alpha})^{(-1)^{p(\alpha)}\,\mathrm{mult}\,\alpha}$. Unfortunately, $ch_{L(\Lambda^0)}$ is not known in general, but in the *boundary level* case, i.e., when $u(k+h^\vee) = h^\vee$, the level of Λ^0 is zero, hence $ch_{L(\Lambda^0)} = 1$, and formula (Ch) gives an explicit expression for $ch_{L(\Lambda)}$. In particular, the character is modular invariant in this case.

Let me mention in conclusion of this section that the Weyl-Kac character formula in the case of 1-dimensional module over an affine Lie algebra $\hat{\mathfrak{g}}$ turns into celebrated Macdonald's identities that express $\hat{R}$ as an infinite series, the special case for $\mathfrak{g} = s\ell_2$ being the Jacobi triple product identity. A sum formula for $\hat{R}$ in the super case is also known [KW3] (see also the talk [Wa1] at the last ICM). In the simplest case of $\mathfrak{g} = s\ell(2|1)$ one gets the identity:

$$\prod_{n=1}^{\infty}\frac{(1-q^n)^2(1-uvq^{n-1})(1-u^{-1}v^{-1}q^n)}{(1-uq^{n-1})(1-u^{-1}q^n)(1-vq^{n-1})(1-v^{-1}q^n)} = \left(\sum_{m,n=0}^{\infty} - \sum_{m,n=-1}^{-\infty}\right)u^mv^nq^{mn}\,,$$

which goes back to Ramanujan and even further back to Kronecker.

4. Quantum reduction for affine Lie superalgebras

4.1. This part of my talk is based on a joint work with S.-S. Roan and M. Wakimoto [KRW], [KW5]. I will explain a general quantum reduction scheme, which is a further development of a number of works. They include [DS], [KS] and [Kh] on classical Drinfeld-Sokolov reduction, and [FF1], [FF2], [FKW], [B], [BT] on its quantization. As in [FF2], the basic idea is to translate the geometric Drinfeld-Sokolov reduction to a homological language as in [KS], and then to quantize this homology complex. Remarkably, this procedure gives, starting from affine superalgebras, a number of very interesting super extensions of the Virasoro algebra and their most interesting representations. I will use the very convenient language of vertex algebras (introduced in [B]), all related notations can be found in [K5].

4.2. Let $\mathfrak{g}$ be a basic simple finite-dimensional Lie superalgebra with a non-degenerate invariant bilinear form $(.\,|\,.)$. Fix a number k and a nilpotent even element f of $\mathfrak{g}$. Include f in an $s\ell_2$-triple $\{e,h,f\}$ so that $[h,e]=2e$, $[h,f]=-2f$, $[e,f]=h$. We have the eigenspace decomposition of $\mathfrak{g}$ with respect to $\operatorname{ad} h$: $\mathfrak{g}=\oplus_{j\in\mathbb{Z}}\mathfrak{g}_j$, and we let $\mathfrak{g}_+=\oplus_{j>0}\mathfrak{g}_j$. The element f defines a non-degenerate skew-supersymmetric bilinear form $\langle .\,,\,.\rangle$ on $\mathfrak{g}_1$ by the formula $\langle a,b\rangle=(f|[a,b])$. Let A_{ne} denote the superspace $\mathfrak{g}_1$ with the form $\langle .\,,\,.\rangle$. Denote by A_{ch} the superspace $\pi\mathfrak{g}_+ + \pi\mathfrak{g}_+^*$, where π stands for the reversal of parity, with the skew-supersymmetric bilinear form defined by $(a,b^*)=b^*(a)$ for $a\in\pi\mathfrak{g}_+, b\in\pi\mathfrak{g}_+^*$, $(\pi\mathfrak{g}_+,\pi\mathfrak{g}_+)=0=(\pi\mathfrak{g}_+^*,\pi\mathfrak{g}_+^*)$.

Let $V^k(\hat{\mathfrak{g}})$ be the universal affine vertex algebra, and let $F^1(A_{ne})$, $F^1(A_{ch})$ be the free fermionic vertex algebras ([K5], § 4.7). Consider the vertex algebra

$$C(\mathfrak{g},f,k)=V^k(\hat{\mathfrak{g}})\otimes F^1(A_{ne})\otimes F^1(A_{ch})\,,$$

and define its charge decomposition $C(\mathfrak{g},f,k)=\oplus_{m\in\mathbb{Z}}C_m$ by letting charge $V^k(\hat{\mathfrak{g}})=$ charge $F^1(A_{ne})=0$, charge $\pi\mathfrak{g}_+=1=-$ charge $\pi\mathfrak{g}_+^*$.

Next, we define a differential d on $C(\mathfrak{g},f,k)$ which makes it a homology complex. For this choose a basis $\{u_i\}_{i\in S'}$ of $\mathfrak{g}_1$ and extend it to a basis $\{u_i\}_{i\in S}$ of $\mathfrak{g}_+$ compatible with its $\mathbb{Z}$-gradation. Denote by $\{\varphi_i\}_{i\in S}$ and $\{\varphi_i^*\}_{i\in S}$ the corresponding dual bases of $\pi\mathfrak{g}_+$ and $\pi\mathfrak{g}_+^*$, and by $\{\Phi_i\}_{i\in S'}$ the corresponding basis of A_{ne}. Consider the following odd field of the vertex algebra $C(\mathfrak{g},f,k)$:

$$\begin{aligned} d(z) \;=\; & \sum_{i\in S}(-1)^{p(u_i)}u_i(z)\otimes\varphi_i^*(z)\otimes 1 \\ & -\tfrac{1}{2}\sum_{i,j,k\in S}(-1)^{p(u_i)p(u_k)}c_{ij}^k\otimes\varphi_k(z)\varphi_i^*(z)\varphi_j^*(z)\otimes 1 \\ & +\sum_{i\in S}(f|u_i)\otimes\varphi_i^*(z)\otimes 1+\sum_{i\in S'}1\otimes\varphi_i^*(z)\otimes\Phi_i(z)\,, \end{aligned}$$

where $[u_i,u_j]=\sum_k c_{ij}^k u_k$ in $\mathfrak{g}_+$, and let $d=\operatorname{Res}_{z=0}d(z)$. (Note that the first two summands of d form the usual differential of a Lie (super)algebra complex.) Then one checks that $[d(z),d(w)]=0$, hence $d^2=0$. It is also clear that $dC_m\subset C_{m-1}$. We define the vertex algebra $W(\mathfrak{g},f,k)$ as the 0^{th} homology of this complex, and call it the *quantum reduction* of the triple $(\mathfrak{g},f,k)$ (actually, it depends only on $\mathfrak{g},k$ and the conjugacy class of f in $\mathfrak{g}_{\bar{0}}$).

One of the fields of the vertex algebra $W(\mathfrak{g}, f, k)$ is the following Virasoro field

$$L(z) = \frac{1}{2(k+h^\vee)} \sum_i :a_i(z)b_i(z): + \frac{1}{2}\partial_z h(z) + \sum_{i\in S}((1-m_i):\partial_z\varphi_i^*(z)\varphi_i(z): - m_i:\varphi_i^*(z)\partial_z\varphi_i(z):) + \tfrac{1}{2}\sum_{i\in S'} g^{ij}:\partial_z\Phi_i(z)\Phi_j(z):,$$

where $\{b_i\}$ and $\{a_i\}$ are dual bases of $\mathfrak{g}$: $(b_i|a_j) = \delta_{ij}$, $[h, u_i] = 2m_i u_i$, (g^{ij}) is the matrix inverse to $(\langle u_i, u_j\rangle)_{i,j\in S'}$. The central charge of $L(z)$ is equal to:

$$c(k) = \frac{k\,\mathrm{sdim}\,\mathfrak{g}}{k+h^\vee} - 3(h|h)k - \sum_{i\in S}(-1)^{p(u_i)}(12m_i^2 - 12m_i + 2) - \tfrac{1}{2}\,\mathrm{sdim}\,\mathfrak{g}_1\,.$$

Here $\mathrm{sdim}\,V = \dim V_{\bar{0}} - \dim V_{\bar{1}}$ is the superdimension of the superspace V. Thus, all $W(\mathfrak{g}, f, k)$ are super-extensions of the Virasoro algebra. Furthermore, for each $\mathrm{ad}\,h$ eigenvector with eigenvalue $-2j$ in the centralizer of f in $\mathfrak{g}$, $W(\mathfrak{g}, f, k)$ contains a field of conformal weight $1+j$ (so that $L(z)$ corresponds to f), and these fields generate the vertex algebra $W(\mathfrak{g}, f, k)$.

Examples. (1) $\mathfrak{g}$ is a simple Lie algebra.
(a) f is a principal nilpotent element. Then $W(\mathfrak{g}, f, k)$ is called the quantum Drinfeld-Sokolov reduction. These algebras and their representations were extensively studied in [FF2], [FB], [FKW] and many other papers. The simplest case of $\mathfrak{g} = s\ell_2$ produces the Virasoro vertex algebra. The case $\mathfrak{g} = s\ell_3$ gives the W_3 algebra [Z].
(b) f is a lowest root vector of $\mathfrak{g}$. These vertex algebras were discussed from a different point of view in [FL] under the name quasi-superconformal algebras. The special case of $\mathfrak{g} = s\ell_3$ was studied from a quantum reduction viewpoint in [Be].
(2) $\mathfrak{g}$ is a simple basic Lie superalgebra and f is an even lowest root vector.
(a) One has the following correspondence:

$\mathfrak{g}$	$spo(2\|1)$	$s\ell(2\|1)$	$s\ell(2\|2)$	$spo(2\|3)$	$D(2,1;a)$
$W(\mathfrak{g}, f, k)$	Neveu-Schwarz	$N=2$	$N=4$	$N=3$	big $N=4$

(In the last two columns one gets an isomorphism after adding one fermion, resp. four fermions and one boson.).
(b) Almost every lowest root vector of a simple component of $\mathfrak{g}_{\bar{0}}$ can be made equal f. This gives all superconformal algebras of [FL] (by definition, they are generated by the Virasoro field, the even fields of weight 1 and N odd fields of weight 3/2), and many new examples.

4.3. Let M be a highest weight module over $\hat{\mathfrak{g}}$. It extends to a vertex algebra module over $V^k(\hat{\mathfrak{g}})$, and we consider the $C(\mathfrak{g}, f, k)$-module $C(M) = M \otimes F^1(A_{ne}) \otimes F^1(A_{ch})$. The element d acts on $C(M)$ and again $d^2 = 0$, hence we can consider homology $H(M) = \oplus_j H_j(M)$ which is a module over $W(\mathfrak{g}, f, k)$. The $W(\mathfrak{g}, f, k)$-module $H(M)$ is called the *quantum reduction* of the $\hat{\mathfrak{g}}$-module M. Using the Euler-Poincaré principle one easily computes the character of $H(M)$ in terms of ch_M. The

basic conjecture of [FKW], [KRW] is that $H(M)$ is irreducible (in particular, at most one $H_j(M)$ is non-zero) if M is admissible.

Examples. (1) $\mathfrak{g} = s\ell_2$. Let k be an admissible level, i.e., k is a rational number with positive denominator u such that $u(k+2) \geq 2$ (recall that $h^\vee = 2$). The set of principal admissible weights of level k is as follows (α is a simple root of $s\ell_2$) [KW1], [K4]:

$$\{\Lambda_{k,j,n} = kD + \tfrac{1}{2}(n - j(k+2))\alpha \,|\, 0 \leq j \leq u-1\,,\, 0 \leq n \leq u(k+2)-2\}\,.$$

Then the quantum reduction of the $\hat{s\ell}_2$-module $L(\Lambda_{k,j,n})$ is the "minimal series" module corresponding to parameters $p = u(k+2)$, $p' = u$ (cf. [BPZ], [K4]):

$$c^{(p,p')} = 1 - 6\frac{(p-p')^2}{pp'}\,,\; h^{(p,p')}_{j+1,n+1} = \frac{(p(j+1) - p'(n+1))^2 - (p-p')^2}{4pp'}\,.$$

The character formula (Ch) for $L(\Lambda_{k,j,n})$ gives immediately all the characters of minimal series.

(2) $\mathfrak{g} = spo(2|1)$. We get all minimal series modules over the Neveu-Schwarz algebra and their characters by quantum reduction of all (not only principal) admissible $\hat{\mathfrak{g}}$-modules.

(3) $\mathfrak{g} = s\ell(2|1)$. Then the boundary admissible levels are $k = m^{-1} - 1$, where $m \in \mathbb{Z}$, $m \geq 2$. One has the following $\hat{w}\hat{\Pi}_m$'s (see Sec. 3.3.), where α_1 and α_2 are odd: $\{\alpha_0 + b_0 K\,,\, \alpha_1 + b_1 K\,,\, \alpha_2 + b_2 K\}$. Here b_i are non-negative integers, $b_0 \geq 1$ and $\sum b_i = m - 1$. The quantum reduction of the corresponding admissible $\hat{\mathfrak{g}}$-modules gives all the minimal series representations of the $N = 2$ superconformal algebra (cf. [FST] and references there). Again, formula (Ch) gives immediately their characters.

(4) $\mathfrak{g} = s\ell(2|2)$ (resp. $spo(2|3)$). In a similar fashion we recover the characters of $N = 4$ [ET] (resp. $N = 3$ [M]) superconformal algebras.

5. Representations of $E(3|6)$ and the standard model

5.1. By a representation of a linearly compact Lie superalgebra L we shall mean a continuous representation in a vector space V with discrete topology (then the contragredient representation is a continuous representation in a linearly compact space V^*). Fix an open subalgebra L_0 of L. We shall assume that V is locally L_0-finite, meaning that any vector of V is contained in a finite-dimensional L_0-invariant subspace (this property actually often implies that V is continuous). These kinds of representations were studied in the Lie algebra case by Rudakov [R]. It is easy to show that such an irreducible L-module V is a quotient of an induced module $\mathrm{Ind}_{L_0}^{L} U = U(L) \otimes_{U(L_0)} U$, where U is a finite-dimensional irreducible L_0-module, by a (unique in good cases) maximal submodule. The induced module $\mathrm{Ind}_{L_0}^{L} U$ is called *degenerate* if it is not irreducible. An irreducible quotient of a degenerate induced module is called a *degenerate* irreducible module.

One of the most important problems of representation theory is to determine all degenerate representations. I will state here the result for $L = E(3|6)$ with $L_0 = \Pi_{j\geq 0}\mathfrak{a}_j$ (see Sec. 1.6.). The finite-dimensional irreducible L_0-modules are actually $\mathfrak{a}_0 = s\ell_3 \oplus s\ell_2 \oplus g\ell_1$-modules (with $\Pi_{j>0}\mathfrak{a}_j$ acting trivially). We shall normalize the generator Y of $g\ell_1$ by the condition that its eigenvalue on $\mathfrak{a}_{-1}$ is $-1/3$. The finite-dimensional irreducible $\mathfrak{a}_0$-modules are labeled by triples $(p,q;r;Y)$, where p,q (resp. r) are labels of the highest weight of an irreducible representation of $s\ell_3$ (resp. $s\ell_2$), so that $p,0$ and $0,q$ label $S^p\mathbb{C}^3$ and $S^q\mathbb{C}^{3*}$ (resp. r labels $S^r\mathbb{C}^2$), and Y is the eigenvalue of the central element Y. Since irreducible $E(3|6)$-modules are unique quotients of induced modules, they can be labeled by the above triples as well.

Theorem 4. [KR1]–[KR3] *The complete list of irreducible degenerate $E(3|6)$-modules consists of four series:* $(p,0;r;-r+\frac{2}{3}p), (p,0;r;r+\frac{2}{3}p+2)$, $(0,q;r;-r-\frac{2}{3}q-2), (0,q;r;r-\frac{2}{3}q)$.

5.2. Remarkably, all four degenerate series occur as cokernels of the differential of a differential complex (M,∇) constructed below (see [KR2] for details). I shall view $E(3|6)$ as a subalgebra of $E(5|10) = \Pi_{j\geq -2}\mathfrak{g}_j$ as in Sect. 1.6., expressed in terms of vector fields and differential forms in the indeterminates x_1, x_2, x_3, $z_+ = x_4$, $z_- = x_5$. Recall that $\mathfrak{g}_0$ is the algenra of divergenceless vector fields with linear coefficients. Let $Y = \frac{2}{3}\sum_i x_i\partial_i - \sum_\epsilon z_\epsilon\partial_\epsilon \in \mathfrak{g}_0$. Here and further $i = 1, 2, 3$, $\epsilon = +,-$, and $\partial_i = \partial/\partial x_i$, $\partial_\epsilon = \partial/\partial z_\epsilon$. Then $\mathfrak{a}_0$ is the centralizer of Y in $\mathfrak{g}_0$ and $\mathfrak{a}_{-1}$ is the span of all elements $d_i^\epsilon = dx_i \wedge dz_\epsilon$.

Consider the following four $\mathfrak{a}_0$-modules (extended to L_0-modules by trivial action of $\mathfrak{a}_j$ with $j>0$):

$$V_I = \mathbb{C}[x_i, z_\epsilon]\,,\ V_{II} = \mathbb{C}[x_i,\partial_\epsilon]_{[2]}\,,\ V_{III} = \mathbb{C}[\partial_i, z_\epsilon]_{[-2]}\,,\ V_{IV} = \mathbb{C}[\partial_i,\partial_\epsilon]\,,$$

where the subscript $[a]$ means that Y is shifted by the scalar a. For each $R = I - IV$ introduce a bigradation $V_R = \oplus_{m,n\in\mathbb{Z}} V_R^{(m,n)}$ by letting $\deg x_i = (1,0)$, $\deg z_\epsilon = (0,1)$, and let $M_R = \mathrm{Ind}_{L_0}^{L} V_R = \oplus_{m,n\in\mathbb{Z}} M_R^{(m,n)}$. Then the non-zero $M_R^{(m,n)}$ are all the degenerate $E(3|6)$-modules of the R^{th} series. We let

$$M = \left(\oplus_{(m,n)\neq(0,0)} M_I^{(m,n)}\right) \oplus M_{II} \oplus M_{III} \oplus \left(\oplus_{(m,n)\neq(0,0)} M_{IV}^{(m,n)}\right)\,.$$

The differentials ∇_k introduced further are elements of $U(L)\otimes \mathrm{End}\ V$ that act on $U(L)\otimes_{U(L_0)} V$ by the formula:

$$(\sum_j u_j\otimes A_j)(u\otimes v) = \sum_j uu_j\otimes A_j v\,.$$

Example. The dual to the ordinary formal de Rham complex is $\mathbb{C}[\partial/\partial x_1,\dots,\partial/\partial x_m]\otimes\Lambda(\partial/\partial\xi_1,\dots,\partial/\partial\xi_m)$ with the differential $d^* = \sum_j \partial/\partial x_j\otimes\xi_j$ and the $\mathbb{Z}$-gradation defined by $\deg\partial/\partial x_i = 0$, $\deg\partial/\partial\xi_i = 1$. Rudakov's theorem [R] says that all irreducible degenerate W_m-modules occur as cokernels of d^*.

Turning now to ∇_k, we let $\Delta^{\pm} = \sum_i d_i^{\pm} \otimes \partial_i$, $\delta_i = d_i^+ \otimes \partial_+ + d_i^- \otimes \partial_-$. Then $\nabla_1 = \Delta^+(1 \otimes \partial_+) + \Delta^-(1 \otimes \partial_-)$ is a well-defined operator on all M_R such that $\nabla_1^2 = 0$. Furthermore there are differentials $\nabla_2 = \Delta^+\Delta^-$, $\nabla_3 = \delta_1\delta_2\delta_3$, ∇_4, ∇_4' and ∇_6 (the explicit expressions of the last three can be found in [KR2]) that sew together these four complexes. These differentials are illustraed by Table M. The white nodes and black marks represent the induced modules of the R^{th} series. The plain arrows represent ∇_1, the dotted arrows represent ∇_2, the interrupted arrows represent ∇_3 and the bold arrows represent ∇_4', ∇_4'' and ∇_6. The white nodes denote the places with zero homology. The black marks denote the places with non-zero homology, also computed in [KR2]. For example, at the star mark the homology is $\mathbb{C}$.

Similar results for $E(3|8)$ and (to a lesser extent) for $E(5|10)$ are given in [KR4].

5.3. The first hint that the Lie superalgebra $E(3|6)$ is somehow related to the Standard Model comes from the observation that its subalgebra $\mathfrak{a}_0$ is isomorphic to the complexified Lie algebra of the group of symmetries of the Standard Model. Table P below lists all $\mathfrak{a}_0$-multiplets of fundamental particles of the Standard Model (see e.g. [O]): the upper part is comprised of three generations of quarks and the middle part of three generations of leptons (these are all fundamental fermions from which matter is built), and the lower part is comprised of the fundamental bosons (which mediate the strong and electro-weak interactions).

multiplets	charges	particles		
$(01, 1, \frac{1}{3})$	$\frac{2}{3}, -\frac{1}{3}$	$\binom{u_L}{d_L}$	$\binom{c_L}{s_L}$	$\binom{t_L}{b_L}$
$(10, 1, -\frac{1}{3})$	$-\frac{2}{3}, \frac{1}{3}$	$\binom{\tilde{u}_R}{\tilde{d}_R}$	$\binom{\tilde{c}_R}{\tilde{s}_R}$	$\binom{\tilde{t}_R}{\tilde{b}_R}$
$(10, 0, -\frac{4}{3})$	$-\frac{2}{3}$	$\tilde{u}_L$	$\tilde{c}_L$	$\tilde{t}_L$
$(01, 0, \frac{4}{3})$	$\frac{2}{3}$	u_R	c_R	t_R
$(01, 0, -\frac{2}{3})$	$-\frac{1}{3}$	d_R	s_R	b_R
$(10, 0, \frac{2}{3})$	$\frac{1}{3}$	$\tilde{d}_L$	$\tilde{s}_L$	$\tilde{b}_L$
$(00, 1, -1)$	$0, -1$	$\binom{\nu_L}{e_L}$	$\binom{\nu_{\mu L}}{\mu_L}$	$\binom{\nu_{\tau L}}{\tau_L}$
$(00, 1, 1)$	$0, 1$	$\binom{\tilde{\nu}_R}{\tilde{e}_R}$	$\binom{\tilde{\nu}_{\mu R}}{\tilde{\mu}_R}$	$\binom{\tilde{\nu}_{\tau R}}{\tilde{\tau}_R}$
$(00, 0, 2)$	1	$\tilde{e}_L$	$\tilde{\mu}_L$	$\tilde{\tau}_L$
$(00, 0, -2)$	-1	e_R	μ_R	τ_R
$(11, 0, 0)$	0	gluons		
$(00, 2, 0)$	$1, -1, 0$	W^+, W^-, Z	(gauge bosons)	
$(00, 0, 0)$	0	γ	(photon)	

Table P

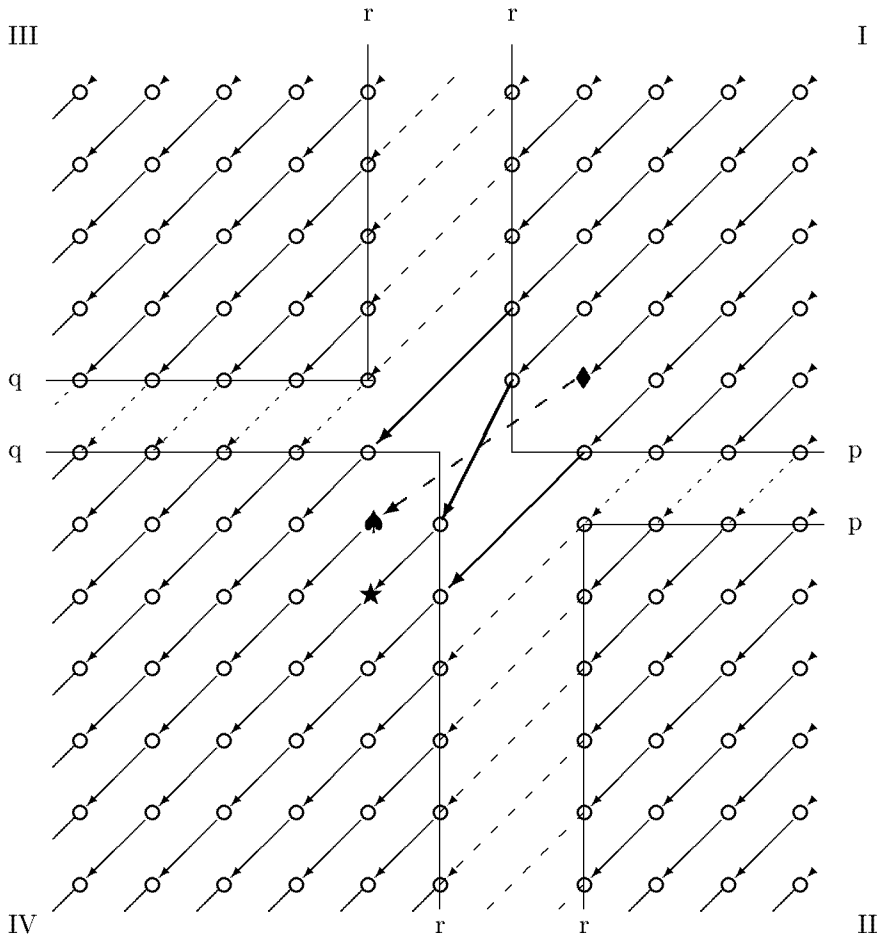


Table M

It is easy to deduce from Theorem 4 that this list of multiplets (plus the multiplets $(11, 0, \pm 2)$) is characterized by the conditions:

(i) $\mathfrak{a}_0$-multiplet occurs in a degenerate irreducible $E(3|6)$-module,
(ii) when restricted to $s\ell_3 \subset \mathfrak{a}_0$, this multiplet contains only 1-dimensional, the two fundamental or the adjoint representation,
(iii) $|Q| \leq 1$ for all particles of the multiplet, where the charge Q of a particle is given by the Gell-Mann-Nishijima formula: $Q = \frac{1}{2}(y + h)$, where y (resp. h) is the Y-eigenvalue (resp. $H = \mathrm{diag}(1, -1) \in s\ell_2$-eigenvalue).

How can we see the number of generations of quarks and leptons? For that order the sequence subcomplexes in Table M by $t = r - q + 3$ in sector IV (time), and replace in them the induced modules by their irreducible quotients. Then we find [KR2] (based on computer calculations by Joris Van der Jeugt) that a fundamental particle multiplet appears in the t^{th} sequence iff $t \geq 1$. Furthermore, for $1 \leq t \leq 7$ we get sequences with various particle contents, but for $t \geq 8$ the particle contents remains unchanged, and it is invariant under the CPT symmetry (though for $t \leq 7$ it is not). The explicit contents is exhibited in [KR2], 659–660.

Remarkably, precisely three generations of leptons occur in the stable region ($t \geq 8$), but the situation with quarks is more complicated: this model predicts a complete fourth generation of quarks and an incomplete fifth generation (with missing down type triplets).

In view of this discussion, it is natural to suggest that the algebra $su_3 + su_2 + u_1$ of internal symetries of the Weinberg-Salam-Glashow Standard Model extends to $E(3|6)$. It is hoped that the representation theory of $E(3|6)$ will shed new light on various features of the Standard Model. I find it quite remarkable that the SU_5 Grand Unified Model of Georgi-Glashow combines the left multiplets of fundamental fermions in precisely the negative part of the consistent gradation of $E(5|10)$ (see Sec. 1.6). This is perhaps an indication of the possibility that the extension from su_5 to $E(5|10)$ algebra of internal symmetries may resolve the difficulties with the proton decay.

References

[ALS] D. Alexseevski, D. Leites and I. Shchepochkina, Examples of simple Lie superalgebras of vector fields, *C.R.Acad.Bul.Sci.* **33** (1980), 1187–1190.

[BPZ] A.A. Belavin, A.M. Polyakov and A.M. Zamolodchikov, Infinite conformal symmetry of critical fluctuations in two dimensions, *J.Stat.Phys.* **34** (1984), 763–774.

[Be] M. Bershadsky, Conformal field theory via Hamiltonian reduction, *Comm. Math. Phys.* **139** (1991), 71–82.

[BT] J. de Boer and T. Tjin, The relation between quantum W algebras and Lie algebras, *Comm. Math. Phys.* **160** (1994), 317–332.

[B] R. Borcherds, Vertex algebras, Kac-Moody algebras, and the Monster, *Proc. Natl. Acad. Sci. USA*, **83** (1986), 3068–3071.

[C] E. Cartan, Les groupes des transformations continués, infinis, simples, *Ann. Sci. Ecole Norm. Sup.* 26 (1909), 93–161.

[CK1] S.-J. Cheng and V.G. Kac, A new $N = 6$ superconformal algebra, *Commun. Math. Phys.* **186** (1997), 219–231.

[CK2] S.-J. Cheng and V.G. Kac, Generalized Spencer cohomology and filtered deformations of $\mathbb{Z}$-graded Lie superalgebras, *Adv. Theor. Math. Phys.* **2** (1998), 1141–1182.

[CK3] S.-J. Cheng and V.G. Kac, Structure of some $\mathbb{Z}$-graded Lie superalgebras of vector fields, *Transformation Groups* **4** (1999), 219–272.

[DK] A. D'Andrea and V.G. Kac, Structure theory of finite conformal algebras, *Selecta Mathematica* **4** (1998), 377–418.

[DS] V.G. Drinfeld and V.V. Sokolov, Lie algebra and the KdV type equations, *Soviet J. Math* **30**, (1985), 1975–2036.

[ET] T. Eguchi and A. Taormina, On the unitary representations of $N = 2$ and $N = 4$ superconformal algebras, *Phys Lett.* **B210** (1988), 125–132.

[FK] D. Fattori and V.G. Kac, Classification of finite simple Lie conformal superalgebras, *J. Algebra* (2002).

[FF1] B.L. Feigin and E.V. Frenkel, Representations of affine Kac-Moody algebras, bozination and resolutions, *Lett. Math. Phys.* **90** (1990), 307–317.

[FF2] B.L. Feigin and E.V. Frenkel, Quantization of Drinfeld-Sokolov reduction, *Phys. Lett.* **B246** (1990), 75–81.

[FST] B.L. Feigin, A.M. Semihkatov, V.A. Sirota and I. Yu Tipunin, Resolutions and characters of irreducible representations of the $N = 2$ superconformal algebra, hep-th/9805179.

[FL] E.S. Fradkin and V. Ya. Linetsky, Classification of superconformal and quasisuperconformal algebras in two dimensions, *Phys. Lett.* **B291** (1992), 71–76.

[FB] E. Frenkel and D. Ben-Zvi, "Vertex algebras and algebraic curves", AMS monographs, **vol. 88**, 2001.

[FKW] E. Frenkel, V. Kac and M. Wakimoto, Characters and fusion rules for W-algebras via quantized Drinfeld-Sokolov reduction, *Comm. Math. Phys.* **147** (1992), 295–328.

[G1] V.W. Guillemin, A Jordan-Hölder decomposition for a certain class of infinite dimensional Lie algebras, *J. Diff. Geom.* **2** (1968), 313–345.

[G2] V.W. Guillemin, Infinite-dimensional primitive Lie algebras, *J. Diff. Geom.* **4** (1970), 257–282.

[GS] V.W. Guillemin and S. Sternberg, An algebraic model of transitive differential geometry, *Bull. Amer. Math. Soc.* **70** (1964), 16–47.

[K1] V.G. Kac, Simple irreducible graded Lie algebras of finite growth, *Math. USSR-Izvestija* **2** (1968), 1271–1311.

[K2] V.G. Kac, Lie superalgebras, *Adv. Math.* **26** (1977), 8–96.

[K3] V.G. Kac, "Infinite-dimensional Lie algebras", Third edition, Cambridge University Press, 1990.

[K4] V.G. Kac, Modular invariance in mathematics and physics, in: "Mathematics into the 21^{st} century", *AMS Centennial Publ.*, 1992, 337–350.

[K5] V.G. Kac, "Vertex algebras for beginners", University Lecture Series, vol. 10 AMS, Providence, RI, 1996. Second edition 1998.

[K6] V.G. Kac, Superconformal algebras and transitive group actions on quadrics, *Comm. Math. Phys.* **186** (1997), 233–252. Erratum, *Comm. Math. Phys.* **217** (2001), 697–698.

[K7] V.G. Kac, Classification of infinite-dimensional simple linearly compact Lie superalgebras, *Adv. Math.* **139** (1998), 1–55.

[K8] V.G. Kac, Classification of infinite-dimensional simple groups of supersymmetries and quantum field theory, "Visions in Mathematics toward 2000", *GAFA*, Special volume (2000), 162–183.

[KL] V.G. Kac and J.W. van de Leur, On classification of superconformal algebras, in: "Strings 88" (Eds S.J. Gates et al.), World Sci., 1989, 77–106.

[KRW] V.G. Kac, S.-S. Roan and M. Wakimoto, Quantum reduction for affine Lie superalgebras, 2002 preprint.

[KR1] V.G. Kac and A.N. Rudakov, Representations of the exceptional Lie superalgebra $E(3,6)$ I: Degeneracy conditions, *Transformation Groups* **7** (2002), 67–86.

[KR2] V.G. Kac and A.N. Rudakov, Representations of the exceptional Lie superalgebra $E(3,6)$ II: Four series of degenerate modules, *Comm. Math. Phys.* **222** (2001), 611–661

[KR3] V.G. Kac and A.N. Rudakov, Representations of the exceptional Lie superalgebra $E(3,6)$ III: Classification of singular vectors.

[KR4] V.G. Kac and A.N. Rudakov, Complexes of modules over exceptional Lie superalgebras $E(3,8)$ and $E(5,10)$, *IMRN* **19** (2002), 1007–1025.

[KW1] V.G. Kac and M. Wakimoto, Modular invariant representations of infinite-dimensional Lie algebras and superalgebras, *Proc. Natl. Acad. Sci* **85** (1989), 4956–4960.

[KW2] V.G. Kac and M. Wakimoto, Classification of modular invariant representations of affine superalgebras, in *Infinite-dimensional Lie algebras and groups*, Adv. Ser. Math. Phys. **vol. 7** World Scientific, 1989, 138–177.

[KW3] V.G. Kac and M. Wakimoto, Integrable highest weight modules over affine superalgebras and number theory, *Progress in Math.* **Vol. 123** (1994), Birkhäuser Boston, 415–456.

[KW4] V.G. Kac and M. Wakimoto, Integrable highest weight modules over affine superalgebras and Appell's function, *Comm. Math. Phys* **215** (2001), 631–682.

[KW5] V.G. Kac and M. Wakimoto, Quantum reduction and representation theory of superconformal algebras.

[Kh] T. Khovanova, *KdV* superequation related to the Lie superalgebra of Neveu-Schwarz-2 string theory, *Teor. Mat. Phys.* **72** (1987), 899–904.

[Ko] Yu. Kochetkoff, Déformations de superalgébres de Buttin et quantification, *C.R. Acad. Sci. Paris* **299**, ser I, no. 14 (1984), 643–645.

[KS] B. Kostant and S. Sternberg, Symplectic reduction, BRS cohomology and infinite-dimensional Clifford algebras, *Ann. Phys.* **176** (1987), 49–113.

[L] S. Lie, Theorie der Transformations gruppen, *Math.Ann.* **16** (1880), 441–528.

[M] K. Miki, The representation theory of the $SO(3)$ invariant superconformal

algebra, *Int. J. Mod. Phys.* **5** (1990), 1293–1318.

[O] L. Okun, "Physics of elementary particles", Nauka, 1988 (in Russian).

[P] A. Polishchuk, M.P. Appell's function and vector bundles of rank 2 on elliptic curves, *Ramanujan J.* **5** (2001), 111–128.

[RS] P. Ramond and J.H. Schwarz, Classification of dual model gauge algebras, *Phys.Lett.B.* **64** (1976), 75–77.

[R] A.N. Rudakov, Irreducible representations of infinite-dimensional Lie algebras of Cartan type, *Math. USSR-Izvestija* **8** (1974), 836–866.

[SV] J.A. Schouten and W. van der Kulk, "Pfaff problem and its generalizations", Clarendon Press, 1949.

[Sh1] I. Shchepochkina, New exceptional simple Lie superalgebras, *C.R. Bul. Sci.* **36**, no. 3 (1983), 313–314.

[Sh2] I. Shchepochkina, Five exceptional simple Lie superalgebras of vector fields, *Funct. Anal. Appl.* **33** (1999), No. 3, 208–219.

[Se] V. Serganova, Characters of irreducible representations of simple Lie superalgebras, *Doc. Math.*, Extra Volume ICM 1998 II, 583–593.

[Wa1] M. Wakimoto, Representation theory of affine superalgebras at the critical level, *Doc. Math.* Extra Volume ICM 1998 II, 605–614.

[Wa2] M. Wakimoto, "Lectures on infinite-dimensional Lie algebras", *World Sci.*, 2001.

[W] B. Yu Weisfeiler, Infinite-dimensional filtered Lie algebras and their connection with graded Lie algebras, *Funct. Anal. Appl.* **2** (1968), 88–89.

[Z] A.B. Zamolodchikov, Infinite additional symmetries in 2-dimensional conformal quantum field theory *Teor. Mat. Phys.* **65** (1985), 1205–1213.

ICM 2002 · Vol. I · 345–362

Some Highlights of Percolation

Harry Kesten*

Abstract

We describe the percolation model and some of the principal results and open problems in percolation theory. We also discuss briefly the spectacular recent progress by Lawler, Schramm, Smirnov and Werner towards understanding the phase transition of percolation (on the triangular lattice).

2000 Mathematics Subject Classification: 60K35, 82B43.
Keywords and Phrases: Percolation, Phase transition, Critical probability, Critical exponents, Power laws, Conformal invariance, SLE.

1. Introduction and description of the percolation model

Percolation was introduced by Broadbent and Hammersley (see [14],[15]) as a probabilistic model for the flow of fluid or a gas through a random medium. It is one of the simplest models which has a phase transition, and is therefore a valuable tool for probabilists and statistical physicists in the study of phase transitions. For many mathematicians percolation on general graphs may be of interest because it exhibits relations between probabilistic and topological properties of graphs. On the applied side, percolation has been used to model the spread of a disease or fire, the spread of rumors or messages, to model the displacement of oil by water, to estimate whether one can build nondefective integrated circuits with certain wiring restrictions.

We shall give a brief survey of some of the important results obtained for this model and list some open problems. The present article is only a very restricted survey and its references (in particular to the physics literature) are far from complete. We apologize to the authors of relevant articles which we have not cited. Earlier surveys are in [42], [21], [22], [37], [38], and the reader can find more elaborate treatments in the books [20], [63] and [55].

*Department of Mathematics, Cornell University, Malott Hall, Ithaca NY 14853, USA. E-mail: kesten@math.cornell.edu

The oldest (indirect) reference to percolation that I know of is a problem submitted to the Amer. Math. Monthly (vol 1, 1894, pp. 211-212) in 1894 by De Volson Wood, Professor of Mechanical Engineering at the Stevens Inst. of Technology in Hoboken NJ. Here is the text of the problem.

"An actual case suggested the following:

An equal number of white and black balls of equal size are thrown into a rectangular box, what is the probability that there will be contiguous contact of white balls from one end of the box to the opposite end ? As a special example, suppose there are 30 balls in the length of the box, 10 in the width and 5 (or 10) layers deep."

Even though percolation theory was not invented to answer this problem, it naturally came to study problems of this kind. By the way, we still have no answer to De Volson Wood's problem. Percolation as a mathematical theory was invented by Broadbent and Hammersley ([14],[15]). Broadbent wanted to model the spread of a gas or fluid through a random medium of small channels which might or might not let gas or fluid pass. To model these channels he took the edges between nearest neighbors on $\mathbb{Z}^d$ and made all edges independently *open* (or passable) with probability p or *closed* (or blocked) with probability $1-p$. Write P_p for the corresponding probability measure on the configurations of open and closed edges (with the obvious σ-algebra generated by the sets determined by the states of finitely many edges). A *path* on $\mathbb{Z}^d$ will be a sequence (finite or infinite) $v_1, v_2, \ldots$ of vertices of $\mathbb{Z}^d$ such that for all $i \geq 1$, v_i and v_{i+1} are adjacent on $\mathbb{Z}^d$. The edges of such a path are the edges $\{v_i, v_{i+1}\}$ between successive vertices and a path is called *open* if all its edges are open. Broadbent's original question amounted to asking for

$$P_p\{\exists \text{ an open path on } \mathbb{Z}^d \text{ form } \mathbf{0} \text{ to } \infty\}. \tag{1.1}$$

This question has an obvious analogue on any infinite connected graph $\mathcal{G}$ with edge set $\mathcal{E}$ and vertex set $\mathcal{V}$. Again one makes all edges independently open or closed with probability p and $1-p$, respectively, and one denotes the corresponding measure on the edge configurations by P_p. E_p is expectation with respect to P_p. An open path is defined as before with $\mathcal{G}$ taking the role of $\mathbb{Z}^d$. A path $(v_1, v_2, \ldots)$ is called *self-avoiding* if $v_i \neq v_j$ for $i \neq j$. (1.1) now is replaced by

$$P_p\{\exists \text{ an infinite self-avoiding open path starting at } v\}, \tag{1.2}$$

with v any vertex in $\mathcal{V}$.

The preceding model is called *bond-percolation.* There is also an analogous model, called *site-percolation.* In the latter model all edges are assumed passable, but the vertices are independently open or closed with probability p or $1-p$, respectively. An open path is now a path all of whose vertices are open. One is still interested in (1.2). Site percolation is more general than bond percolation in the sense that the positivity of (1.2) for some v in bond-percolation on a graph $\mathcal{G}$ is equivalent to the positivity of (1.2) for some v in site-percolation on the covering graph or line graph of $\mathcal{G}$. However, site percolation on a graph may not be equivalent to bond percolation on another graph (see [40], Section 2.5 and Proposition 3.1).

Unless otherwise stated we restrict ourselves in the remaining sections to site percolation. We shall often use $\mathcal{V}$ and $\mathcal{E}$ to denote the vertex and edge set of whatever graph we are discussing at that moment, without formally introducing the graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. It should be clear from the context what $\mathcal{V}$ and $\mathcal{E}$ stand for in such cases. For $A \subset \mathcal{V}$, we shall use $|A|$ to denote the number of vertices in A. Further if A, B and C are sets of vertices, then $A \leftrightarrow B$ means that there exists an open path from some vertex in A to some vertex in B, while $A \overset{C}{\leftrightarrow} B$ means that there exists an open path with all its vertices in C, from some vertex in A to some vertex in B. In particular, with some abuse of notation, we have

$$\{|\mathcal{C}(v)| = \infty\} = \{v \leftrightarrow \infty\}.$$

Definition 1 *We call a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ quasi-transitive if there is a finite set of vertices V_0, such that for each vertex v there is a graph automorphism of $\mathcal{G}$ which maps v to one of the vertices in V_0.*

All vertices which can be mapped by a graph automorphism to a fixed $v_0 \in V_0$ are equivalent for our purposes. In a quasi-transitive graph each vertex is equivalent to one of finitely many vertices. A special subclass is formed by the *transitive graphs*, which have $|V_0| = 1$, so that all vertices are equivalent for our purposes. (For example, the Cayley graph of a finitely generated group is transitive.)

We shall restrict ourselves here to graphs which are

$$\text{connected, infinite but locally finite, and quasi-transitive.} \tag{1.3}$$

Graphs which satisfy (1.3) automatically have countable vertex sets and edgesets. We define, for $v \in \mathcal{V}$,

$$\begin{aligned} \theta^v(p) &= P_p\{v \leftrightarrow \infty\} \\ &= P_p\{\exists \text{ an infinite self-avoiding open path starting at } v\}. \end{aligned} \tag{1.4}$$

For a quasi-transitive graph $\theta^v(p) = \theta^{v_0}(p)$ for some $v_0 \in V_0$. It is an easy consequence of the FKG inequality that either $\theta^v(p) > 0$ for all v or $\theta^v(p) = 0$ for all v (see [40], Section 4.1). We call $\theta^v(p)$ the *percolation probability (from v)*. Much of the earlier work on percolation theory deals with properties of the function $p \mapsto \theta^v(p)$, or more generally with the full distribution of the so-called cluster sizes. The *cluster* $\mathcal{C}(v)$ of the vertex v is the set of all points which are connected to the origin by an open path. By convention, this always contains the vertex v itself (even if v itself is closed in the case of site percolation). The clusters are the maximal components of the graph with vertex set $\mathcal{V}$ and with an edge between two sites only if they are adjacent on $\mathcal{G}$ and are both open. $\theta^v(p)$ is just the P_p-probability that $|\mathcal{C}(v)| = \infty$.

2. Existence of phase transition and related properties of the critical probability

The most important property of the percolation model is that it exhibits a *phase transition*, that is, there exists a threshold value p_c such that the global behavior of the system is quite different in the two regions $p < p_c$ and $p > p_c$. To make this more precise let us consider the percolation probability as a function of p. It is non-decreasing. This is easiest seen from Hammersley's ([31]) joint construction of percolation systems for all $p \in [0,1]$ on $\mathcal{G}$. Let $\{U(v), v \in \mathcal{V}\}$ be independent uniform $[0,1]$ random variables. Declare v to be p-open if $U(v) \leq p$. Then the configuration of p-open vertices has distribution P_p for each $p \in [0,1]$. Clearly the collection of p-open vertices is nondecreasing in p and hence also $\theta(\cdot)$ is nondecreasing. Clearly $\theta^v(0) = 0$ and $\theta^v(1) = 1$. Roughly speaking the graph of $\theta^v(\cdot)$ (for a fixed v) therefore looks as in figure 1, but not all the features exhibited in this figure have been proven.

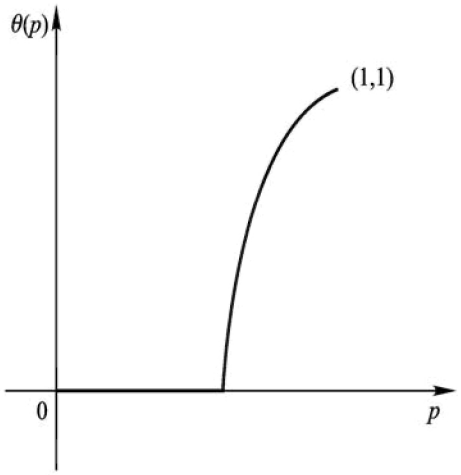


Figure 1: Graph of θ. Many aspects of this graph are still conjectural.

The *critical probability* is defined as

$$p_c = p_c(\mathcal{G}) = \sup\{p : \theta^v(p) = 0\}. \tag{2.1}$$

As remarked after (1.4) this is independent of v. By definition we then have

$$P_p\{|\mathcal{C}(v)| = \infty\} = 0 \text{ for } p < p_c, v \in \mathcal{V},$$

so that

$$\text{all clusters are finite a.s. } [P_p] \text{ when } p < p_c. \tag{2.2}$$

On the other hand, for $p > p_c$ there is a strictly positive P_p-probability that $|\mathcal{C}(v)|$ is infinite. It then follows from Kolmogorov's zero-one law that

$$P_p\{\text{some } |\mathcal{C}(v)| = \infty\} = 1, \quad p > p_c. \tag{2.3}$$

Thus the global behavior of the system is quite different for $0 \leq p < p_c$ and for $p_c < p < 1$. We therefore can say that there is a phase transition at p_c, provided

the intervals $[0, p_c)$ and $(p_c, 1]$ are both nonempty. It is easy to see from a so-called Peierls argument (just as in [29]) that $p_c(\mathcal{G}) > 0$ for any graph $\mathcal{G}$ of bounded degree (and hence certainly if (1.3) holds). It is much harder to do show that $p_c(\mathcal{G}) < 1$ holds for certain $\mathcal{G}$. Hammersley [30] proved this for bond-percolation on $\mathbb{Z}^d$, but a similar argument works for site-percolation and various other periodic graphs. (Basically, we say that $\mathcal{G}$ is *periodic* or can be *periodically imbedded in* $\mathbb{R}^d$ if $\mathcal{V}$ can be imbedded in $\mathbb{R}^d$ (with $d \geq 2$) such that $\mathcal{V}$ as well as the edges of $\mathcal{G}$, as represented by the straight line segments between the pairs of vertices adjacent in $\mathcal{G}$, form a subset of $\mathbb{R}^d$ which is invariant under translations by d linearly independent vectors. If this is the case we call d the dimension of $\mathcal{G}$. We refer the reader to [40], Section 2.1 for details.) Thus

Theorem 1

$$0 < p_c(\mathbb{Z}^d) < 1.$$

Thus, at least on $\mathbb{Z}^d$, there really is a phase transition. On any graph one says that the system is in the *subcritical (supercritical) phase* if $p < p_c$ (respectively, $p > p_c$). Because percolation is such a simple model with a phase transition, percolation has received a great deal of attention from physicists. Percolation is one of the Potts models, corresponding to the parameter q in the Potts model equal to 1; the famous Ising model for magnetism is essentially the same as the Potts model with $q = 2$. One hopes that understanding of the percolation model will help understand all the Potts models and even the more general Fortuin-Kasteleyn or random cluster models (see [23], which also explains the relation, due to Fortuin and Kasteleyn, between random cluster models and Potts models).

The exact value of $p_c(\mathcal{G})$ is known only for a handful of graphs, and all of these are periodic two-dimensional graphs. This leads to

Open problem 1: Find $p_c(\mathcal{G})$ for a wide class of graphs.

However, it is generally agreed that the solution to this problem would not have any explanatory value. The critical probabilities which have been determined so far depend heavily on special symmetry properties of the underlying graph, and the values of these critical probabilities vary with the graph. One has therefore moved on to properties which are believed to be shared by large classes of graphs; see Section 4 below. The rigorously known critical probabilities can be found in [38], Chapter 3. Here we merely mention the one case which will be important later on:

$$p_c(\text{site percolation on triangular lattice}) = \frac{1}{2}. \tag{2.4}$$

Also known is the following asymptotic result, both for the site and for the bond version:

$$p_c(\mathbb{Z}^d) \sim \frac{1}{2d} \text{ as } d \to \infty. \tag{2.5}$$

This has been proven by several people; [35] gives the best higher order terms in (2.5).

One can define another critical probability as the threshhold value for the finiteness of the clustersize of a fixed vertex. Thus,

$$p_T(\mathcal{G}) = \sup\{p : E_p\{|\mathcal{C}(v)|\} = \infty\}. \tag{2.6}$$

Since $P_p\{|\mathcal{C}(v)| = \infty\} > 0$ for $p > p_c$, it is obvious that $E_p\{|\mathcal{C}(v)|\} = \infty$ for all $p > p_c$, so that $p_T(\mathcal{G}) \le p_c(\mathcal{G})$. It was a crucial step in establishing the known values for p_c to show that $p_T(\mathcal{G}) = p_c(\mathcal{G})$. The original proof of this fact was only for bond percolation on $\mathbb{Z}^2$ ([39]; this proof made strong use of crossing probabilities similar to those appearing in De Volson Wood's problem in Section 1). Proofs of $p_T = p_c$ for some other special lattices are in [65] and [40]). Later Menshikov ([51]) and Aizenman and Barsky ([1]) gave independent and different proofs of exponential decay of the distribution of $|\mathcal{C}(v)|$ for $p < p_c$. This is a cornerstone of the subject and is of course a much stronger statement than $p_T = p_c$.

Theorem 2 *(Menshikov and Aizenman and Barsky) Assume that $\mathcal{G}$ is periodic. Then for $p < p_c(\mathcal{G})$ there exists constants $0 < C_1, C_2 < \infty$ such that*

$$P_p\{|\mathcal{C}(v)| \ge n\} \le C_1 e^{-C_2 n}, \quad n \ge 0. \tag{2.7}$$

(2.7) gives a basic estimate for the subcritical phase. By an earlier "subadditivity" argument of [45] (2.7) can be sharpened to a "local limit theorem" (see [20], Theorem 6.78): for each $p < p_c$ there exists a $0 < C_3(p) < \infty$ such that

$$\lim_{n\to\infty} -\frac{1}{n} \log P_p\{|C(v)| = n\} = C_3(p). \tag{2.8}$$

These results give us a measure of control over the subcritical phase. In the supercritical phase many estimates rely on another fundamental result of percolation theory, which was proven by Grimmett and Marstrand [24]. The simplest form of the result is as follows:

Theorem 3

$$p_c(\mathbb{Z}^d) = \lim_{k\to\infty} p_c(\mathbb{Z}_+^2 \times \{1, 2, \ldots, k\}^{d-2}). \tag{2.9}$$

One may replace $\mathbb{Z}_+^2$ by $\mathbb{Z}^2$ here.

The graph appearing in the right hand side here consists of a finite number of copies of the first quadrant in $\mathbb{Z}^2$ or of the whole $\mathbb{Z}^2$. Thus (before the limit is taken) this graph looks very much like $\mathbb{Z}^2$ and many of the special tools for percolation on $\mathbb{Z}^2$ can be applied to this graph. Because of this one could prove a number of results on $\mathbb{Z}^d$ for $p > \lim_{k\to\infty} p_c(\mathbb{Z}_+^2 \times \{1, 2, \ldots, k\}^{d-2})$. Theorem 3 now shows that these results hold throughout the supercritical regime (at least when $\mathcal{G} = \mathbb{Z}^d$ or a similar graph). As an example of this situation we mention a result of [43], namely the right hand inequality in (2.10) (the left hand inequality is due to [3]): For site percolation on $\mathbb{Z}^d$ with $p > p_c(\mathbb{Z}^d)$ there exist $0 < C_4(p), C_5(p) < \infty$ such that

$$C_4(p) \le -\frac{1}{n^{(d-1)/d}} \log P_p\{|C(v)| = n\} \le C_5 \tag{2.10}$$

for all large n.

Open problem 2: Does

$$\lim_{n\to\infty} -n^{-(d-1)/d} \log P_p\{|C(v)| = n\} \text{ exist ?}$$

(under the conditions for (2.10)). The reader should notice the contrast between (2.6), (2.7) — which give exponential decay for the clustersize distribution in the subcritical case — and (2.10) which corresponds to a "stretched exponential" for the tail of the clustersize in the supercritical case. The tail of this distribution at criticality, i.e., for $p = p_c$ will be discussed in Section 4.

3. Uniqueness of infinite clusters and properties of the percolation probability

It is natural to ask "how many infinite clusters can there be ?" In [52] it is shown that for periodic graphs for each p, exactly one of the following three situations prevails:
$P_p\{\text{there is no infinite open cluster}\} = 1$,
$P_p\{\text{there is exactly one infinite open cluster}\} = 1$ or
$P_p\{\text{there are infinitely many infinite open clusters}\} = 1$.
As pointed out in [58], the proof of [52] carries over to any quasi-transitive graph by a zero-one law for events which are invariant under graph automorphisms. Of course, the first alternative here holds for $p < p_c$, but can the last situation occur for some $p \geq p_c$? The first proof that this is impossible on $\mathbb{Z}^d$ is in [4]. This proof was improved and generalized a few times, but the most elegant, and by now standard, proof is due to Burton and Keane [16]. Their method works for any amenable graph. To make this precise we define for any set $W \subset \mathcal{V}$,

$$\partial W = \{w \in \mathcal{V} : w \notin W \text{ but } w \text{ is adjacent to some } v \in W\}.$$

We call the graph $\mathcal{G}$ *amenable* if there exists a sequence $\{W_n\} \subset \mathcal{V}$ for which $|\partial W_n|/|W_n| \to 0$.

Theorem 4 *(Burton and Keane) If $\mathcal{G}$ satisfies* (1.3)*, and if $\mathcal{G}$ is amenable, then for all $p \in [0,1]$*

$$P_p\{\textit{there exist more than one infinite open cluster}\} = 0. \tag{3.1}$$

The proof of this result is the same as in [16], except that one should argue on the *expected number of encounter points* where Burton and Keane use the ergodic theorem to make the number of encounter points itself large. (We owe this observation to O. Häggström.) Simple examples (such as a regular tree) show that (3.1) does not have to hold for nonamenable graphs. This is one example of a relation between percolation properties and algebraic/topological properties of the underlying graph (see [10], [50] and [8] and some of their references for other examples). What can be said about uniqueness/nonuniqueness in the nonamenable case ? Benjamini and Schramm [10] introduced a further critical probability:

$$p_u = p_u(\mathcal{G}) := \inf\{p : \text{a.s. } [P_p] \text{ there is a unique infinite cluster}\}. \tag{3.2}$$

By definition $p_u \geq p_c$. We have $p_c < p_u = 1$ on a regular b-ary tree (in which all vertices have degree $b+1$) with $b \geq 2$. The first example of a graph

with $p_c < p_u < 1$ was given in [25]. Note that there is no a priori reason why uniqueness should be monotone in p, that is why uniqueness a.s. $[P_{p'}]$ should imply uniqueness a.s. $[P_{p''}]$ whenever $p'' \geq p'$. This has been proven to be the case for graphs satisfying (1.3). More precisely, the following theorem (and somewhat more) is proven in [57] (see also [26] and [27]):

Theorem 5 *Let $\mathcal{G}$ satisfy* (1.3) *and let the percolation configurations on $\mathcal{G}$ be constructed simultaneously for all $p \in [0,1]$ by Hammersley's method described in the beginning of Section 2. Let $N(p)$ be the number of p-open infinite clusters. Then a.s.,*

$$N(p) = \begin{cases} 0 & \text{for } p \in [0, p_c) \\ \infty & \text{for } p \in (p_c, p_u) \\ 1 & \text{for } p \in (p_u, 1]. \end{cases}$$

Note that this theorem does not give the value of $N(p)$ at $p = p_c$ or p_u (see also the lines after Theorem 1.2 in [26] and Open problem 3 below).

Other obvious questions concern the smoothness of the function $\theta^v(\cdot)$, and in particular whether this function is continuous. Clearly $p \mapsto \theta^v(p)$ is always continuous for $p < p_c$, since $\theta^v(p) = 0$ for all such p. Russo [53] noted that $\theta^v(\cdot)$ is everywhere right continuous and [11] proved that (under (1.3)) if for some fixed $p_0 > p_c$ there is a.s. $[P_{p_0}]$ a unique infinite cluster, then $\theta^v(\cdot)$ is also left continuous at p_0. Thus, under (1.3), the remaining problem is

Open problem 3: Is $p \mapsto \theta^v(p)$ (left) continuous on $[p_c, p_u]$?

On $\mathbb{Z}^d$ which has $p_c(\mathbb{Z}^d) = p_u(\mathbb{Z}^d)$, continuity is equivalent to $\theta(p_c) = 0$. It has long been conjectured that this is the case. It is known that this holds for $d \geq 19$ by the theory of Hara and Slade [34]; actually this deals with bond percolation, but should go through also for site percolation on $\mathbb{Z}^d$. It also follows from work of Harris [36] and the author [40], Theorem 3.1, that continuity holds when $d = 2$ (both for bond and site percolation). [8] and [9] prove that on a Cayley graph of a non-amenable group there is no percolation at p_c.

4. Behavior at and near p_c

From now on we shall restrict ourselves to transitive graphs which are periodically imbedded in $\mathbb{R}^d$, so that the origin is a vertex of the graph. Since all vertices are equivalent in a transitive graph, we drop the superscript v from various quantities such as $\theta(p)$; we further write $\mathcal{C}$ for the open cluster of the origin.

We saw in (2.7) and (2.10) that the probability of a cluster of size $n < \infty$ decays exponentially or as a stretched exponential in the subcritical and supercritical regime, respectively. The behavior at criticality is quite different. In fact, it is believed that there exists constants $0 < C_i < \infty$ such that

$$C_6 n^{-(d-1)/2} \leq P_{p_c}\{|C(v)| \geq n\} \leq C_7 n^{-C_8}. \tag{4.1}$$

Indeed, for periodic graphs in dimension $d = 2$ the left hand inequality is proven in [12], but the argument remains valid in any dimension. In an Abelian sense

one knows even more (see the proof of Proposition 10.29 in [20]). The right hand inequality of (4.1) for certain two-dimensional graphs can be found in [40], Theorem 8.2, while for $d \geq 19$ with $\delta = 2$ it follows from [6] and [33]. It is natural to conjecture that

$$P_{p_c}\{|C(v)| \geq n\} \approx n^{-1/\delta} \tag{4.2}$$

for some $\delta = \delta(\mathcal{G}) > 0$, where $a(n) \approx b(n)$ means $\log(a(n)/\log b(n) \to 1$ as $n \to \infty$. (One may conjecture that (4.2) and similar relations below hold with an even stronger interpretation of $\approx$ but we shall not pursue this here.)

(4.2) is one example of a so-called *power law.* Another (conjectured) power law is for $P_{p_c}\{v' \leftrightarrow v''\}$. It is believed that for some constant η

$$P_{p_c}\{v' \leftrightarrow v''\} \approx |v' - v''|^{2-d-\eta} \text{ as } |v' - v''| \to \infty. \tag{4.3}$$

Here $|v|$ denotes the ℓ^1 norm of the image of v under the imbedding into $\mathbb{R}^d$. Again this is supported by the following partial result for periodic graphs whose image under the periodic imbedding into $\mathbb{R}^d$ is invariant under permutations of the coordinates. For such graphs (4.1) implies that there exist constants $0 < C_i < \infty$ such that

$$C_9|v' - v''|^{-C_{10}} \leq P_{p_c}\{v' \leftrightarrow v''\} \leq C_{11}|v' - v''|^{-C_{12}}. \tag{4.4}$$

Physicists also conjectured that various quantities behave like powers of $|p - p_c|$ as $p \to p_c, p \neq p_c$. Such conjectures are analogues of results which were known (often on a nonrigorous basis) or conjectured for related models. They were also rigorously known for quite some time for percolation on regular trees. In addition, Hara and Slade in a series of important papers (see in particular [33], [34]) have developed the so-called lace expansion technique to give us a good understanding of percolation in high dimensions. Roughly speaking, they prove many of the physicists conjectures for bond percolation on $\mathbb{Z}^d$ with $d \geq 19$, by showing that most quantities show mean field behavior near p_c, that is, they have the same singularity on $\mathbb{Z}^d$ with $d \geq 19$ as on a regular tree. It is believed that this will remain true for $d > 6$. In fact Hara and Slade can prove their results for percolation in any dimension > 6 for what they call "spread out" models. (These have $\mathbb{Z}^d$ as vertex set but there may be some open bonds between points which are not nearest neighbors on $\mathbb{Z}^d$.)

The most common of the conjectured power laws (with the traditional names for the exponents) are as follows. Here $A(p) \approx B(p)$ means $\log A(p)/\log B(p) \to 1$ as $p \to p_c$ (with $p \neq p_c$).

$$\left(\frac{d}{dp}\right)^3 \Big[\sum_{n=1}^{\infty} \frac{1}{n} P_p\{|\mathcal{C}| = n\}\Big] = \left(\frac{d}{dp}\right)^3 E_p\{|\mathcal{C}|^{-1}\} \approx |p - p_c|^{-1-\alpha}, \tag{4.5}$$

$$\theta(p) \approx (p - p_c)^{\beta}, \quad p \downarrow p_c, \tag{4.6}$$

$$\chi(p) := E\{|\mathcal{C}(v)|; |\mathcal{C}(v)| < \infty\} \approx |p - p_c|^{-\gamma}. \tag{4.7}$$

Another power law is supposed to hold for the so-called *correlation length,* $\xi(p)$. Intuitively speaking, if $p \neq p_c$, the correlation length is the minimal size a cube should have so that one can detect from a typical percolation configuration in such

a cube that p is not equal to p_c. On scales which are small with respect to the correlation length, the system is expected to behave as if it is critical. On the other hand, on scales which are large with respect to the correlation length, one should be able to partition the system into cubes of edgelength equal to a large multiple of the correlation length and regard these cubes as "supersites"; for $p > p_c$ (respectively $p < p_c$) these supersites should behave as sites in site percolation with a p value close to 1 (respectively close to 0). On such scales the details of the lattice other than its dimension should play little role, if any. Several possible formal definitions are in use for the correlation length. Here we define the correlation length $\xi(p)$ by

$$[\xi(p)]^{-1} = \lim_{n\to\infty} -\frac{1}{n} \log P_p\{\mathbf{0} \leftrightarrow ne_1, \mathbf{0} \not\leftrightarrow \infty\}, \tag{4.8}$$

where e_1 is the first coordinate vector. Strictly speaking, [19] only proves that this is a good definition for (bond or site) percolation on $\mathbb{Z}^d$, but this definition should make sense with minor changes for percolation on general periodic graphs. The conjectured powerlaw then takes the form

$$\xi(p) \approx |p - p_c|^{-\nu}. \tag{4.9}$$

Other power laws have been conjectured for electrical conductance and for the graph-theoretical length of an open crossing between opposite faces of a cube.

In all these cases proofs of power *bounds* instead of actual power laws are known for many graphs which are periodically imbedded in $\mathbb{R}^d$ with $d = 2$ or d large (see [40], Chapter 8, [20], Chapter 10, [33], [34]). For instance,

$$C_{13}|p - p_c|^{-1} \le \chi(p) \le C_{14}|p - p_c|^{-C_{15}} \text{ for } p < p_c. \tag{4.10}$$

In fact, the left hand inequality holds for all d (see [5], [20], Theorem 10.28).

Most remarkable is the conjecture of "universality". That is, it is generally believed that each of the so-called critical exponents $\alpha, \beta, \gamma, \delta, \eta, \nu$ depends for periodic graphs on the dimension d only, and not on the details of the graph $\mathcal{G}$. For instance, they should have the same value for bond and for site percolation on $\mathbb{Z}^2$ and on the triangular lattice. This is in contrast to the critical probablity p_c, which definitely *does* depend on the details of $\mathcal{G}$. For this reason the principal concern these days is to establish power laws and universality, and little attention is being paid to open problem 1. (See next section for more on what is now known.)

There are also nonrigorous arguments to derive simple relations between various of these exponents. These are the so-called scaling laws:

$$\alpha + \beta(\delta + 1) = 2 \tag{4.11}$$

$$\gamma + 2\beta = \beta(\delta + 1) \tag{4.12}$$

$$\gamma = \nu(2 - \eta) \tag{4.13}$$

$$d\nu = \gamma + 2\beta \text{ for } 2 \le d \le 6. \tag{4.14}$$

The last relation, which involves the dimension d is called a hyper-scaling law. These scaling relations, except (4.11), have been established for many graphs with

$d = 2$ (assuming that the exponents exist; see [41]) and are also known for high d. There is also a "conditional proof" of the hyperscaling relation . That is, (4.14) (or rather the relation $(2 - \eta) = d(\delta - 1)/(\delta + 1)$) has been shown to be implied by other (not yet established) laws for percolation ([13]). It is widely believed that these other laws hold for $3 \leq d \leq 6$.

There are predictions by physicists of the values of these exponents when $d = 2$ or when d is large. In fact, as we already pointed out, it is believed that all these exponents are even independent of d for $d > 6$. The existence of these exponents and their predicted values have now been proven to be correct when $\mathcal{G}$ is the triangular lattice ([60], [61], [62], [48]). It is also known that these exponents (except for α and perhaps η) exist and take the same values as on a regular tree for $d \geq 19$ ([33], [34], [32]).

This section raises the obvious and very extensive

Open problem 4: Prove power laws, universality and scaling relations.

In the next section we shall describe some of the progress made on this problem in dimension 2. We already mentioned the work of Hara and Slade in high dimensions. No progress has been made in dimensions 3, 4 and 5. So we may pose a more modest problem for these dimensions.

Open problem 5: Find upper and lower power bounds for $\theta(p), \chi(p)$, $\xi(p)$ and $P_{p_c}\{|\mathcal{C}| \geq n\}$ when $3 \leq d \leq 5$.

As we pointed out above, bounds on one side are already known for most of these quantities, but as far as we know no bound of the form

$$\xi(p) \leq C_{16}|p - p_c|^{-C_{17}} \tag{4.15}$$

has been proven for $3 \leq d \leq 5$, not even for $p < p_c$. This is probably the most fundamental bound to prove, from which several other bounds might follow. Note that it is not hard to see that on $\mathbb{Z}^d$

$$\xi(p) \geq C_{18}|\log(p_c - p)|^{-(d-1)/d}(p_c - p)^{-1/d} \text{ for } p < p_c. \tag{4.16}$$

Indeed, the proof of the left hand inequality in (4.4) actually gives

$$P_{p_c}\{0 \overset{[0,n]^d}{\leftrightarrow} ne_1\} \geq C_9 n^{-3(d-1)}. \tag{4.17}$$

From this one trivially has for $p < p_c$

$$\begin{aligned} P_p\{\mathbf{0} \leftrightarrow kne_1\} &\geq \left[P_p\{\mathbf{0} \overset{[0,n]^d}{\leftrightarrow} ne_1\}\right]^k \\ &\geq \left[\left(\frac{p}{p_c}\right)^{(n+1)^d} P_{p_c}\{\mathbf{0} \overset{[0,n]^d}{\leftrightarrow} ne_1\}\right]^k \\ &\geq \left[\left(\frac{p}{p_c}\right)^{(n+1)^d} C_9 n^{-3(d-1)}\right]^k. \end{aligned} \tag{4.18}$$

Now take $n = \left[|\log(p_c - p)|\right]^{1/d}(p_c - p)^{-1/d}$ and estimate $\xi(p)$ from

$$\begin{aligned} [\xi(p)]^{-1} &= \lim_{k\to\infty} -\frac{1}{kn}\log P_p\{\mathbf{0} \leftrightarrow kne_1, \mathbf{0} \nleftrightarrow \infty\} \\ &= \lim_{k\to\infty} -\frac{1}{kn}\log P_p\{\mathbf{0} \leftrightarrow kne_1\} \text{ for } p < p_c. \end{aligned}$$

A somewhat different aspect of the behavior of critical percolation concerns the random variable

$$N(v) := \inf\{\text{number of closed vertices in any path from } \mathbf{0} \text{ to } v\}. \tag{4.19}$$

If no percolation occurs for $p = p_c$, then $N(v) \to \infty$ as $v \to \infty$, a.s. $[P_{p_c}]$. For bond percolation on $\mathbb{Z}^2$ it is known that $[\sigma_{p_c}(N(v))]^{-1}[N(v) - E_{p_c}N(v)]$ satisfies a central limit theorem with $E_{p_c}N(v) \asymp \log|v|$ and $\sigma_{p_c}(N(v))$, the standard deviation of $N(v)$, of order $[\log|v|]^{1/2}$ (see [44]). On $\mathbb{Z}^d$ with $d \geq 3$ it is only known ([18]) that $N(v) = O(|v|^{\varepsilon})$ a.s. $[P_{p_c}]$, for every $\varepsilon > 0$.
Open problem 6: Improve the bound for $N(v)$ and find a limit theorem for $N(v)$ in dimension ≥ 3.

5. Conformal invariance and SLE

In this section we only consider graphs which are periodically imbedded in $\mathbb{R}^2$. Special attention will be paid to site percolation on the triangular lattice.

We already briefly discussed the interpretation of the correlation length in the preceding section. In view of (4.4), the definition (4.8) assigns the value ∞ to the correlation length when $p = p_c$, at least if $P_{p_c}\{|\mathcal{C}| = \infty\} = 0$, as is widely believed (and is known for $d = 2$ or $d \geq 19$). Thus the correlation length is not a useful length scale for critical percolation. Other than the spacing between vertices, there seems to be no lengthscale which plays a role for critical percolation. In this case one may hope to take some sort of limit without normalization of a critical percolation system in a larger and larger region. It is not clear in what topology one should take a limit. Matters look somewhat friendlier if one fixes a region and considers a limit as the spacing between vertices tends to zero. Even then it is not clear what topology will be most useful for taking a limit. A discussion of these issues can be found in [2] and the beginning of [59]. Putting this problem aside, let us first ask for limits of simple quantities such as crossing probabilities. Let D be a Jordan domain in $\mathbb{R}^2$ with a smooth boundary and let A_1 and A_2 be two disjoint arcs of ∂D. Identify $\mathcal{G}$ with its periodic imbedding in $\mathbb{R}^2$. (This imbedding is not unique, but for the present purposes we can just fix some imbedding.) We can then define $\delta\mathcal{G}$ as the result of multiplying the image of $\mathcal{G}$ under the imbedding by a factor $\delta > 0$. This image has vertices located at $\{\delta v : v \in \mathcal{V}\}$ and edges between two points $\delta v', \delta v''$ if and only if v' and v'' are adjacent in $\mathcal{G}$. For any percolation configuration on $\mathcal{G}$ we say that there exists an open path on $\delta\mathcal{G}$ from A_1 to A_2 in D if there is an open path $v_1, \ldots, v_m$ on $\mathcal{G}$ such that $\delta v_i \in D$ for $2 \leq i \leq m-1$ and the edge between δv_1 and δv_2 intersects A_1 and the edge between δv_{m-1} and δv_m intersects A_2. We then define

$$h(D, A_1, A_2, \delta) := P_{p_c}\{\exists \text{ open path on } \delta\mathcal{G} \text{ from } A_1 \text{ to } A_2 \text{ in } D\}, \tag{5.1}$$

and ask whether this has a limit as $\delta \downarrow 0$. (Here is were contact is made with De Volson Wood's problem in the Amer. Math. Monthly.) It is conjectured that this limit, call it $h(D, A_1, A_2)$, exists, and moreover that it is conformally invariant. By

this we mean that if ϕ is a conformal map from D onto $D' = \phi(D)$ which extends to a homeomorphism between $\overline{D}$:= closure of D and $\overline{D}'$, then

$$h(D, A_1, A_2) = h(\phi(D), \phi(A_1), \phi(A_2)). \tag{5.2}$$

Conformal invariance of a limit of critical percolation had been conjectured by physicists (see [17] and its references) on the grounds that this had been found in related models. The stress on studying this for crossing probabilities is due to [46], which also credits Aizenman with the formulation of conformal invariance for crossing probabilities (actually in a slightly more general form than (5.2)). Cardy used conformal invariance and the Riemann mapping theorem to equate $h(D, A_1, A_2)$ to $h(\mathbb{H}, [z, 0], [1, \infty))$, where $\mathbb{H}$ is the upper half plane and $z \in (-\infty, 0)$ a suitable point on the boundary of $\mathbb{H}$, i.e., the real axis. He then derived (nonrigorously) a differential equation for $h(\mathbb{H}, [z, 0], [1, \infty))$ as a function of z. From this he obtained an explicit formula for $h(\mathbb{H}, [z, 0], [1, \infty))$, and hence for $h(D, A_1, A_2)$ in special cases, such as when D is a rectangle and A_1, A_2 two opposite sides of D. In an astonishing paper Smirnov [60] succeeded in showing that for site percolation on the triangular lattice, the limit $h(D, A_1, A_2)$ indeed exists and is conformally invariant. To do this Smirnov introduces an extra variable $z \in \overline{D}$, and considers

$$\begin{aligned} f(z, D, A_1, A_2, \delta) := P_{p_c}\{&\exists \text{ self-avoiding open path on } \delta\mathcal{G} \text{ from } A_1 \\ &\text{to } B_1 \cup A_2 \text{ in } D \text{ which separates } z \text{ from } B_2\}, \end{aligned}$$

where B_1, B_2 are the arcs on ∂D between A_1 and A_2 (i.e., the boundary of D consists of the four arcs A_1, B_1, A_2, B_2 and one successively traverses these arcs as one goes around the boundary of D in one direction). He now shows that any limit of $f(z, D, A_1, A_2, \delta)$ along a subsequence $\delta_n \downarrow 0$ is a harmonic function of $z \in D$ which has to satisfy certain boundary conditions which uniquely determine the limit. Therefore $\lim_{\delta \downarrow 0} f(z, D, A_1, A_2, \delta_n)$ exists. Moreover, the limit is conformally invariant, because it is characterized as the harmonic function which satisfies a certain boundary condition. The original problem for the crossing probabilities $h(D, A_1, A_2, \delta)$ can be treated as a special case, by letting z approach the single point in $\overline{A_2} \cap \overline{B_1}$. One can find the limit function $h(D, A_1, A_2)$ explicitly if D is a rectangle, and A_1, B_1, A_2, B_2 its sides, and thereby one can recover Cardy's formula.

Somewhat before Smirnov, Schramm [59] had introduced *stochastic Loewner evolutions* (SLE) in order to describe a scaling limit of growing random sets (and in particular the scaling limit of loop erased random walk in dimension two, and related processes). For percolation, the simplest version of SLE is probably the so-called chordal SLE (see [54]), described as follows. Let $\mathbb{H}$ and $\overline{\mathbb{H}}$ be the open upper and closed upper half plane, respectively, and let $\{B(t)\}_{t \geq 0}$ be a standard Brownian motion starting at 0. Let $g_t(z)$ be the solution of the Loewner equation

$$\frac{\partial g_t(z)}{\partial t} = \frac{2}{g_t(z) - \xi(t)}, \quad g_0(z) = z \tag{5.3}$$

with $\xi(t) = \sqrt{\kappa} B(t)$ for some parameter $\kappa > 0$. The solution to (5.3) exists for $t < \tau(z) := \inf\{s : 0 \text{ is a limit point of the set } \{g_u(z) - \xi(u), u < s\}\}$. Define

$H_t := \{z \in \mathbb{H} : \tau(z) > t\}, K_t = \{z \in \overline{H} : \tau(z) \leq t\}$. Chordal SLE_κ is the collection of maps $\{g_t : t \geq 0\}$. It turns out that g_t is the unique conformal homeomorphism from H_t onto $\mathbb{H}$ for which $\lim_{z\to\infty}[g_t(z) - z] = 0$. It is shown in [54] that for $\kappa \neq 8$ there exists a continuous path $\gamma : [0, \infty) \to \overline{\mathbb{H}}$ such that K_t is the hull of $\gamma[0, t]$, that is, K_t is the closure of the union of the bounded components of $\overline{\mathbb{H}} \setminus \gamma[0, t]$. In many situations one can also start with the path γ and then define g_t as the conformal homeomorphism from its hull K_t onto $\mathbb{H}$. This must then satisfy a Loewner equation (5.3). γ is called the *trace* of the corresponding SLE process.

Schramm ([59]) showed that the scaling limit of loop erased random walk can be described by an analogue of SLE_2 in the unit disc. ([59] still had to assume that this scaling limit exists and is conformally invariant, but this has since been proven in [49]). In [59] Schramm expresses the belief that SLE_6 is appropriate for the description of the scaling limit of the boundary of percolation clusters. This has been proven to be correct for percolation on the triangular lattice.

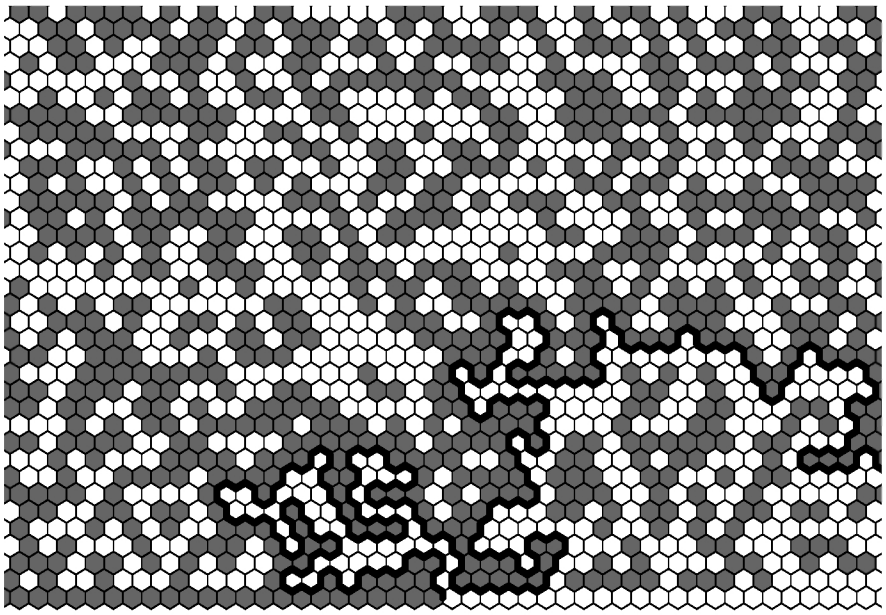


Figure 2: The exploration process, which separates the open (white) hexagons from the closed (black) ones. We thank Oded Schramm for providing us with this figure.

To give a specific example, consider the hexagonal lattice, imbedded in $\mathbb{R}^2$ in such a way that the hexagonal faces which intersect the x-axis have their centers on this axis and that the origin lies on the common boundary of two such faces. Make the hexagonal faces in the upper half plane independently open or closed with probability 1/2. If one thinks of the centers of the hexagonal faces as vertices of the triangular lattice then one sees that this is equivalent to critical site percolation

on the triangular lattice on $\mathbb{H}$ (recall that its critical probability equals 1/2). Now impose the boundary condition that all faces with center on the positive (negative) x-axis are open (closed, respectively). There is then a curve,γ_δ, in the upper half plane and running on the boundaries of some of the hexagonal faces, which starts at 0 and traverses the boundary between the open cluster of the positive x-axis and the closed cluster of the negative x-axis. This curve is called the *exploration process*; see Figure 2. The distribution of this curve γ_δ converges to the distribution of the trace of SLE_6 (as the mesh size goes to zero, and using the Hausdorff metric on the space of curves, determined up to parametrization) (see [60], [61]). Actually these references discuss the analogous situation on an equilateral triangle instead of $\mathbb{H}$ and concentrate on showing the existence of the limit. The identification of the limit as SLE_6 is based on the work of Lawler, Schramm and Werner ([47], [64]).

To prove this result Smirnov ([60], [61]) first uses a compactness argument to show that any sequence $\delta_n \downarrow 0$ has a subsequence along which the distribution of γ_{δ_n} converges to some distribution $\mu^{e\cdot p\cdot}$ on Hölder continuous curves. Then he proves that $\mu^{e\cdot p\cdot}$ is independent of the subsequence $\{\delta_n\}$ by showing that $\mu^{e\cdot p\cdot}$ has certain properties which characterize SLE_6. This of course also shows that $\mu^{e\cdot p\cdot}$ is the distribution of the trace of SLE_6. The second step relies on a reduction of various $\mu^{e\cdot p\cdot}$-probabilities to crossing probabilities of the form (5.1) and on the existence and conformal invariance of the limit of (5.1). In addition it relies on a "locality property." Note that one can construct γ_δ from "local" information only; at any step γ_δ turns to the right (left) if its tip has a closed (respectively, open) hexagon in front of it.

SLE turns out to be the perfect tool for calculating critical exponents. Lawler, Schramm, Smirnow and Werner in [48] and [62] were able to use the correspondence with SLE_6 to prove for percolation on the triangular lattice, not only Cardy's formula, but also the power laws (4.2), (4.3), (4.6), (4.7) and (4.9) with the values for $\nu, \beta, \gamma, \delta$ and η which were predicted by physicists (see [62] for relevant references). It was further shown in [7] by Beffara that the Hausdorff dimension of the trace of SLE_6 is 7/4. Thus, this is also the Hausdorff dimension of the exploration process "in the scaling limit." This dimension had already been predicted by Saleur and Duplantier [56].

References

[1] M. Aizenman and D. J. Barsky, Sharpness of the phase transition in percolation models, Comm. Math. Phys. **108** (1987) 489–526.

[2] M. Aizenman and A. Burchard, Hölder regularity and dimension bounds for random curves, Duke Math. J. **99** (1999) 419–453.

[3] M. Aizenman, F. Delyon and B. Souillard, Lower bounds on the cluster size distribution, J. Stat. Phys. **23** (1980) 267–280.

[4] M. Aizenman, H. Kesten and C. M. Newman, Uniqueness of the infinite cluster and continuity of connectivity functions for short- and long-range percolation, Comm. Math. Phys. **111** (1987) 505–532.

[5] M. Aizenman and C. M. Newman, Tree graph inequalities and critical behavior in percolation models, J. Statist. Phys. **36** (1984) 107–143.
[6] D. J. Barsky and M. Aizenman, Percolation critical exponents under the triangle condition, Ann. Probab. **19** (1991) 1520–1536.
[7] V. Beffara, Hausdorff dimensions for SLE_6, preprint 2002, arXiv:mathPR/0204208.
[8] I. Benjamini, R. Lyons, Y. Peres and O. Schramm, Group-invariant percolation on graphs, Geom. Funct. Anal. **9** (1999) 29–66.
[9] I. Benjamini, R. Lyons, Y. Peres and O. Schramm, Critical percolation on any non-amenable group has no infinite clusters, Ann. Probab. **27** (1999) 1347–1356.
[10] I. Benjamini and O. Schramm, Percolation beyond $\mathbb{Z}^d$, many questions and few answers, Electr. Comm. Probab. **1** (1996) 71–82.
[11] J. van den Berg and M. Keane, On the continuity of the percolation probability function, 61–65 in Particle Systems, Random Media and Large Deviations (R. Durrett ed.), Contemporary Math. 26, Amer. Math. Soc., 1984.
[12] J. van den Berg and H. Kesten, Inequalities with applications to percolation and reliability, J. Appl. Probab. **22** (1985) 556–569.
[13] C. Borgs, J. T. Chayes, H. Kesten and J. Spencer, Uniform boundedness of critical crossing probabilities implies hyperscaling, Random Structures Algorithms **15** (1999) 368–413.
[14] S. R. Broadbent, Contribution to Discussion on symposium on Monte Carlo methods, J. Roy. Statist. Soc. B **16** (1954) 68.
[15] S. R. Broadbent and J. M. Hammersley, Percolation processes I. Crystals and mazes, Proc. Cambr. Phil. Soc. **53** (1957) 629–641.
[16] R. M. Burton and M. Keane, Density and uniqueness in percolation, Comm. Math. Phys. **121** (1989) 501–505.
[17] J. Cardy, Lectures on conformal invariance and peroclation, arXiv:math-ph/0103018v2.
[18] L. Chayes, On the critical behavior of the first passage time in $d \geq 3$, Helvetica Physica Acta **64** (1991) 1–16.
[19] J. T. Chayes, L. Chayes, G. R. Grimmett, H. Kesten and R. H. Schonmann, The correlation length for the high-density phase of Bernoulli percolation, Ann. Probab. **17** (1989) 1277–1302.
[20] G. R. Grimmett, Percolation, Springer-Verlag, 1989.
[21] G. R. Grimmett, Percolation and disordered systems, 153–300 in Ecole d'Eté de Probabilités de Saint Flour XXVI-1996 (P. Bernard ed.), Lecture Notes in Math, vol. 1665, Springer, 1997.
[22] G. R. Grimmett, Percolation, 547–575 in Development of Mathematics 1950–2000 (Jean-Paul Pier ed.), Birkhäuser, 2000.
[23] G. R. Grimmett, The random cluster model, article in Encyclopedia of Math. Sciences, to appear.
[24] G. R. Grimmett and J. M. Marstrand, The supercritical phase of percolation is well behaved, Proc. Roy. Soc. (London) Ser. A **430** (1990) 439–457.
[25] G. R. Grimmett and C. M. Newman, Percolation in $\infty + 1$ dimensions, 167–

190 in Disorder in Physical Systems (G. R. Grimmett and D. J. A. Welsh eds.) Clarendon Press, 1990.
[26] O. Häggström and Y. Peres, Monotonicity of uniqueness for percolation on Cayley graphs: all infinite clusters are born simultaneously, Probab. Theory Rel. Fields **113** (1999) 271–285.
[27] O. Häggström, Y. Peres and R. H. Schonmann, Percolation on transitive graphs as a coalescent process: relentless merging followed by simultaneous uniqueness, 69–90 in Perplexing Problems in Probability (M. Bramson and R. Durrett eds.), Progress in Probability 44, Birkhäuser Boston, 1999.
[28] J. M. Hammersley, Percolation processes II. The connective constant, Proc. Cambr. Phil. Soc. **53** (1957) 642–645.
[29] J. M. Hammersley, Percolation processes. Lower bounds for the critical probability, Ann. Math. Statist. **28** (1957) 790–795.
[30] J. M. Hammersley, Bornes supérieures de la probabilité critique dans un processus de filtration, 17–37 in Le Calcul de Probabilités et ses Applications, CNRS, Paris, 1959.
[31] J. M. Hammersley, A Monte Carlo solution of percolation in a cubic lattice, 281–298 in Methods in Computational Physics, vol. I (B. Alder, S. Fernbach, M. Rotenberg eds.), Academic Press, 1963.
[32] T. Hara, R. van der Hofstad and G. Slade, Critical two-point functions and the lace expansion for spread-out high-dimensional percolation and related models, preprint, 2001.
[33] T. Hara and G. Slade, Mean-field critical behavior for percolation in high dimensions, Comm. Math. Phys. **128** (1990) 333–391.
[34] T. Hara and G. Slade, Mean-field behavior and the lace expansion, 87–122 in Probability and Phase Transition (G. R. Grimmett ed.), Kluwer, 1994.
[35] T. Hara and G. Slade, The self-avoiding walk and percolation critical points in high dimensions, Combin. Probab. Comput. **4** (1995) 197–215.
[36] T. E. Harris, A lower bound for the critical probability in a certain percolation process, Proc. Cambr. Phil. Soc. **56** (1960) 13–20.
[37] S. Havlin and D. Ben-Avraham, Diffusion in disordered media, Adv. in Phys. **36** (1987) 695–798.
[38] B. D. Hughes, Random Walks and Random Environments, vol. 2, Oxford Univ. Press, 1996.
[39] H. Kesten, The critical probability of bond percolation on the square lattice equals 1/2, Comm. Math. Phys. **74** (1980) 41–59.
[40] H. Kesten, Percolation Theory for Mathematicians, Birkhäuser, 1982.
[41] H. Kesten, Scaling relations for 2D-percolation, Comm. Math. Phys. **109** (1987) 109–156.
[42] H. Kesten, Percolation theory and first-passage percolation, Ann. Probab. **15** (1987) 1231–1271.
[43] H. Kesten and Y. Zhang, The probability of a large finite cluster in supercritical Bernoulli percolation, Ann. Probab. **18** (1990) 537–555.
[44] H. Kesten and Y. Zhang, A central limit theorem for "critical" first- passage percolation in two dimensions, Probab. Theory Rel. Fields **107** (1997) 137–

160.
[45] H. Kunz and B. Souillard, Essential singularity in percolation problems and asymptotic behavior of cluster size distribution, J. Stat. Phys. **19** (1978) 77–106.
[46] R. Langlands, P. Pouliot and Y. Saint-Aubin, Conformal invariance in two-dimensional percolation, Bull. Amer. Math. Soc. (N.S.) **30** (1994) 1–61.
[47] G. F. Lawler, O. Schramm and W. Werner, Values of Brownian intersection exponents I:Half-plane exponents, Acta Math. **187**(2001)237–273.
[48] G. F. Lawler, O. Schramm and W. Werner, One-arm exponent for critical 2D percolation, preprint 2001, arXiv:math.PR/0108211.
[49] G. F. Lawler, O. Schramm and W. Werner, Conformal invariance of planar loop-erased random walks and uniform spanning trees, preprint 2002, arXiv:math.PR/0112234.
[50] R. Lyons and Y. Peres, Probability on Trees and Networks, available at http://php.indiana.edu/rdlyons/prbtree/prbtree.html, 1997.
[51] M. V. Menshikov, Coincidence of critical points in percolation problems, Soviet Math. Doklady **33** (1986) 856–859.
[52] C. M. Newman and L. S. Schulman, Infinite clusters in percolation models, J. Statist. Phys. **26** (1981) 613–628.
[53] L. Russo, A note on percolation, Z. Wahrsch. verw. Gebiete **43** (1978) 39–48.
[54] S. Rohde and O. Schramm, Basic properties of SLE, preprint 2001, arXiv:math PR/0106036.
[55] M. Sahimi, Applications of Percolation Theory, Taylor & Francis, 1994.
[56] H. Saleur and B. Duplantier, Exact determination of the percolation hull exponent in two dimensions, Phys. Review Letters **58** (1987) 2325–2328.
[57] R. H. Schonmann, Stability of infinite clusters in supercritical percolation, Probab. Theory Rel. Fields **113** (1999) 287–300.
[58] R. H. Schonmann, Percolation in $\infty + 1$ dimensions at the uniqueness threshold, 53–67 in Perplexing Problems in Probability (M. Bramson and R. Durrett eds.), Progress in Probability 44, Birkhäuser Boston, 1999.
[59] O. Schramm, Scaling limits of loop-erased random walks and uniform spanning trees, Isr. J. Math. **118** (2000) 221–228.
[60] S. Smirnov, Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits, C. R. Acad. Sci. Paris **333** (2001) 239–244.
[61] S. Smirnov, in preparation.
[62] S. Smirnov and W. Werner, Critical exponents for two-dimensional percolation, Math. Res. Letters **8** (2001) 729–744.
[63] D. Stauffer and A. Aharony, Introduction to Percolation Theory, second ed., Taylor and Francis, 1991.
[64] W. Werner, Critical exponents, conformal invariance and planar Brownian motion, 87–103 in Proc. 3rd Europ. Congress Math., Prog. Math. Vol 202, Birkhäuser, 2001.
[65] J. C. Wierman, Bond percolation on honeycomb and triangular lattices, Adv. Appl. Probab. **13** (1981) 293–313.

ICM 2002 · Vol. I · 363–382

Cohomology of Moduli Spaces

Frances Kirwan*

Abstract

Some recent progress towards understanding the cohomology of moduli spaces of curves is described. Madsen and Weiss have announced a proof of a generalisation of Mumford's conjecture on the stable cohomology of these moduli spaces $\mathcal{M}_g$, and other contributors have made advances related to Faber's conjectures concerning the tautological ring of $\mathcal{M}_g$.

2000 Mathematics Subject Classification: 14H10, 14H15, 32G15, 55R40, 57R20.
Keywords and Phrases: Moduli spaces of curves, Mapping class groups, Tautological cohomology classes, Harer stabilisation.

Moduli spaces arise in classification problems in algebraic geometry (and other areas of geometry) when, as is typically the case, there are not enough discrete invariants to classify objects up to isomorphism. In the case of nonsingular complex projective curves (or compact Riemann surfaces) the genus g is a discrete invariant which classifies the curve regarded as a topological surface, but does not determine its complex structure when $g > 0$. For each $g \geq 0$ there is a moduli space $\mathcal{M}_g$ whose points correspond bijectively to isomorphism classes of nonsingular complex projective curves of genus g, and whose geometric structure reflects the way such curves can vary in families depending on parameters. The topology of these moduli spaces $\mathcal{M}_g$ and their compactifications has been studied for several decades, and important progress has been made recently on some long-standing questions concerning their cohomology.

In his fundamental paper [93] Mumford considered some tautological cohomological classes $\kappa_j \in H^{2j}(\mathcal{M}_g)$ for $j = 1, 2, \ldots$ which extend naturally to the Deligne-Mumford compactification $\overline{\mathcal{M}}_g$. Much work on the cohomology of $\mathcal{M}_g$ has concentrated on its tautological ring, which is the subalgebra of its rational cohomology ring (or of its Chow ring) generated by these tautological classes.

One reason for the importance of the tautological ring of $\mathcal{M}_g$ is its relationship with the stable cohomology ring $H^*(\mathcal{M}_\infty; \mathbb{Q})$. It was proved by Harer [47] that the cohomology $H^k(\mathcal{M}_g; \mathbb{Q})$ of $\mathcal{M}_g$ in degree k is independent of the genus g when

*Oxford University, UK. E-mail: kirwan@maths.ox.ac.uk

$g \gg k$, making it possible to define $H^k(\mathcal{M}_\infty;\mathbb{Q})$ as $H^k(\mathcal{M}_g;\mathbb{Q})$ for suitably large g. Mumford conjectured that the stable cohomology ring $H^*(\mathcal{M}_\infty;\mathbb{Q})$ is freely generated by the tautological classes $\kappa_1, \kappa_2, \ldots$ and Miller [83] and Morita [85] proved part of this conjecture by showing that the natural map

$$\mathbb{Q}[\kappa_1, \kappa_2, \ldots] \to H^*(\mathcal{M}_\infty;\mathbb{Q})$$

is injective. The remainder of Mumford's conjecture, that this map is surjective, remained unproved for nearly two decades. However Madsen and Tillmann [80] found an interpretation of Mumford's map on the level of homotopy, which they conjectured should be a homotopy equivalence. Very recently a proof of their conjecture, using h-principle arguments combined with Harer stabilisation, has been announced by Madsen and Weiss [81, 103], and from this Mumford's conjecture follows.

The tautological ring of $\mathcal{M}_g$ for finite g has many beautiful properties. Faber [26] conjectured that when $g \geq 2$ the tautological ring of $\mathcal{M}_g$ looks like the algebraic cohomology ring of a nonsingular complex projective variety of dimension $g-2$, and that it is generated by the tautological classes $\kappa_1, \kappa_2, \ldots, \kappa_{[g/3]}$ with no relations in degrees at most $[g/3]$. He also provided an explicit conjecture for a complete set of relations among these generators. Progress has been made by many contributors towards Faber's conjectures, and also related problems on moduli spaces linked to $\mathcal{M}_g$. In particular Morita [90, 91] has recently proved that the rational cohomological version of the tautological ring of $\mathcal{M}_g$ is indeed generated by $\kappa_1, \kappa_2, \ldots, \kappa_{[g/3]}$. The definition of the tautological ring has also been extended to the compactification $\overline{\mathcal{M}}_{g,n}$ of the moduli space $\mathcal{M}_{g,n}$ of nonsingular curves of genus g with n marked points (motivated by Witten's conjectures [107], proved by Kontsevich [71], on intersection pairings on $\overline{\mathcal{M}}_{g,n}$).

The moduli spaces $\mathcal{M}_g$ and $\mathcal{M}_{g,n}$ have other younger and more sophisticated relatives, such as the moduli spaces $\mathcal{M}_{g,n}(X,\beta)$ which parametrise holomomorphic maps $f : \Sigma \to X$ from a nonsingular complex projective curve Σ of genus g with n marked points satisfying $f_*[\Sigma] = \beta \in H_2(X)$, and their compactifications $\overline{\mathcal{M}}_{g,n}(X,\beta)$ which parametrise 'stable' maps. Intersection theory on $\overline{\mathcal{M}}_{g,n}(X,\beta)$ is fundamental to Gromov-Witten theory and quantum cohomology for X, with numerous applications in the last decade to enumerative geometry. The Virasoro conjecture of Eguchi, Hori and Xiong provides relations among the descendent Gromov-Witten invariants of X, and its recent proof by Givental [39] for $X = \mathbb{P}^n$ implies part of Faber's conjecture by [37].

Other relatives of $\mathcal{M}_g$ include the moduli spaces of pairs (Σ, E) where Σ is a nonsingular curve and E is a stable vector bundle over Σ, and their compactifications; intersection theory on these relates intersection theory on $\overline{\mathcal{M}}_g$ and intersection theory on moduli spaces of bundles over a fixed curve, which is by now quite well understood.

1. Moduli spaces of curves

The study of algebraic curves, and how they vary in families, has been fundamental to algebraic geometry since the beginning of the subject, and has made huge advances in the last few decades [3, 52]. The concept of moduli as parameters describing as efficiently as possible the variation of geometric objects was initiated in Riemann's famous paper [99] of 1857, in which he observed that an isomorphism class of compact Riemann surfaces of genus $g \geq 2$ 'hängt ... von $3g-3$ stetig veränderlichen Grössen ab, welche die Moduln dieser Klasse genannt werden sollen'. In modern terminology, Riemann's observation is the statement that the dimension of $\mathcal{M}_g$ is $3g-3$ if $g \geq 2$. It was not until the 1960s that precise definitions and methods of constructing moduli spaces were given by Mumford in [92] following ideas of Grothendieck. Roughly speaking, the moduli space $\mathcal{M}_g$ is the set of isomorphism classes of nonsingular complex projective curves[1] of genus g, endowed with the structure of a complex variety in such a way that any family of nonsingular complex projective curves parametrised by a base space S induces a morphism from S to $\mathcal{M}_g$ which associates to each $s \in S$ the isomorphism class of the curve parametrised by s. The moduli spaces $\mathcal{M}_g$ can be constructed in several different ways, including

- as orbit spaces for group actions,
- via period maps and Torelli's theorem, and
- using Teichmüller theory.

The first of these is a standard method for constructing many different moduli spaces, using Mumford's geometric invariant theory [92, 95, 105] or more recent ideas due to Kollár [70] and to Mori and Keel [64]. Geometric invariant theory provides a beautiful compactification of $\mathcal{M}_g$ known as the Deligne-Mumford compactification $\overline{\mathcal{M}}_g$ [15]. This compactification is itself modular: it is the moduli space of (Deligne-Mumford) stable curves (i.e. complex projective curves with only nodal singularities and finitely many automorphisms). $\overline{\mathcal{M}}_g$ is singular but in a relatively mild way; it is the quotient of a nonsingular variety by a finite group action [77].

The moduli space $\mathcal{M}_{g,n}$ of nonsingular complex projective curves of genus g with n marked points has a similar compactification $\overline{\mathcal{M}}_{g,n}$ which is the moduli space of complex projective curves with n marked nonsingular points and with only nodal singularities and finitely many automorphisms. Finiteness of the automorphism group of such a curve Σ is equivalent to the requirement that any irreducible component of genus 0 (respectively 1) has at least 3 (respectively 1) special points, where 'special' means either marked or singular in Σ (and the condition on genus 1 components here is redundant when $g \geq 2$).

The second method of construction using the period matrices of curves leads to a different compactification $\tilde{\mathcal{M}}_g$ of $\mathcal{M}_g$ known as the Satake (or Satake-Baily-Borel) compactification. Like the Deligne-Mumford compactification, $\tilde{\mathcal{M}}_g$ is a complex projective variety, but the boundary $\tilde{\mathcal{M}}_g \setminus \mathcal{M}_g$ of $\mathcal{M}_g$ in $\tilde{\mathcal{M}}_g$ has (complex) codimension 2 for $g \geq 3$ whereas the boundary $\Delta = \overline{\mathcal{M}}_g \setminus \mathcal{M}_g$ of $\mathcal{M}_g$ in $\overline{\mathcal{M}}_g$ has codimension 1. Each of the irreducible components $\Delta_0, \ldots, \Delta_{[g/2]}$ of Δ is the closure of a locus of curves with exactly one node (irreducible curves with one node

[1] All complex curves and real surfaces will be assumed to be connected.

in the case of Δ_0, and in the case of any other Δ_i the union of two nonsingular curves of genus i and $g-i$ meeting at a single point). The divisors Δ_i meet transversely in $\overline{\mathcal{M}}_g$, and their intersections define a natural decomposition of Δ into connected strata which parametrise stable curves of a fixed topological type. The boundary of $\mathcal{M}_{g,n}$ in $\overline{\mathcal{M}}_{g,n}$ has a similar description, but now as well as the genus of each irreducible component it is necessary to keep track of which marked points it contains.

The third method of constructing $\mathcal{M}_g$, via Teichmüller theory, leaves algebraic geometry altogether.

2. Teichmüller theory and mapping class groups

Important recent advances concerning the cohomology of the moduli spaces $\mathcal{M}_g$ (in particular [80, 81, 90, 91, 103]) have been proved by topologists via the link between these moduli spaces and mapping class groups of compact surfaces.

Let us fix a compact oriented smooth surface Σ_g of genus $g \geq 2$, and let $\mathrm{Diff}_+\Sigma_g$ be the group of orientation preserving diffeomorphisms of Σ_g. Then the mapping class group Γ_g of Σ_g is the group

$$\Gamma_g = \pi_0(\mathrm{Diff}_+\Sigma_g)$$

of connected components of $\mathrm{Diff}_+\Sigma_g$. It acts properly and discontinuously on the Teichmüller space $\mathcal{T}_g$ of Σ_g, which is the space of conformal structures on Σ_g up to isotopy. The Teichmüller space $\mathcal{T}_g$ is homeomorphic to $\mathbb{R}^{6g-6}$, and its quotient by the action of the mapping class group Γ_g can be identified naturally with the moduli space $\mathcal{M}_g$. This means that there is a natural isomorphism of rational cohomology

$$H^*(\mathcal{M}_g;\mathbb{Q}) \cong H^*(\Gamma_g;\mathbb{Q}). \tag{2.1}$$

The corresponding integral cohomology groups are not in general isomorphic because of the existence of nonsingular complex projective curves with nontrivial automorphisms. If, however, we work with the moduli spaces $\mathcal{M}_{g,n}$ of nonsingular complex projective curves of genus g with n marked points, then when n is large enough such marked curves have no nontrivial automorphisms (cf. [52] p 37) and

$$H^*(\mathcal{M}_{g,n};\mathbb{Z}) \cong H^*(\Gamma_{g,n};\mathbb{Z})$$

where $\Gamma_{g,n}$ is the group of connected components of the group $\mathrm{Diff}_+\Sigma_{g,n}$ of orientation preserving diffeomorphisms of Σ_g which fix n chosen points on Σ_g.

In fact [20] the components of $\mathrm{Diff}_+\Sigma_g$ are contractible when $g \geq 2$, so there is also a natural isomorphism

$$H^*(\Gamma_g;\mathbb{Z}) \cong H^*(B\mathrm{Diff}_+\Sigma_g;\mathbb{Z})$$

where $B\mathrm{Diff}_+\Sigma_g$ is the universal classifying space for $\mathrm{Diff}_+\Sigma_g$. This means that any cohomology class of the mapping class group Γ_g can be regarded as a characteristic class of oriented surface bundles, while any rational cohomology class of Γ_g can be regarded as a rational cohomology class of the moduli space $\mathcal{M}_g$.

The mapping class group Γ_g can be described in a group theoretical way. Γ_g acts faithfully by outer automorphisms (that is, the action is defined modulo inner automorphisms) on the fundamental group $\pi_1(\Sigma_g)$ of Σ_g, which is generated by $2g$ elements $a_1, \ldots, a_g, b_1, \ldots, b_g$ subject to one relation $\prod_j a_j b_j a_j^{-1} b_j^{-1} = 1$, and the image of Γ_g in $\mathrm{Out}(\pi_1(\Sigma_g))$ is the group of outer automorphisms of $\pi_1(\Sigma_g)$ which act trivially on $H_2(\pi_1(\Sigma_g); \mathbb{Z})$ [110]. Γ_g has a finite presentation [106] with generators represented by Dehn twists (diffeomorphisms of Σ_g obtained by cutting Σ_g along a regularly embedded circle, twisting one of the resulting boundary circles through 2π and reglueing). There are similar descriptions of $\Gamma_{g,n}$ [110, 33].

3. Stable cohomology

Harer [47] proved in the 1980s that $H^k(\Gamma_g; \mathbb{Z})$ and $H^k(\Gamma_{g+1}; \mathbb{Z})$ are isomorphic when $g \geq 3k-1$, and the same is true for $\Gamma_{g,n}$ and $\Gamma_{g+1,n}$. This bound was improved by Ivanov [55] and Harer [50] made a further improvement. Since $H^*(\Gamma_g; \mathbb{Q})$ is isomorphic to $H^*(\mathcal{M}_g; \mathbb{Q})$, this means that the rational cohomology group $H^k(\mathcal{M}_g; \mathbb{Q})$ is independent of g for $g \gg k$, and we can define the stable cohomology ring

$$H^*(\mathcal{M}_\infty; \mathbb{Q})$$

so that $H^k(\mathcal{M}_\infty; \mathbb{Q}) \cong H^k(\mathcal{M}_g; \mathbb{Q})$ for $g \gg k$.

Harer's stabilisation map can be defined as follows. We choose a smooth identification of Σ_{g+1} with a connected sum of a smooth surface Σ_g of genus g and a surface Σ_1 of genus 1 (and if we have marked points we make sure they all correspond to points in Σ_g). Let Γ_{g+1,Σ_1} be the subgroup of Γ_{g+1} consisting of mapping classes represented by diffeomorphisms from Σ_{g+1} to itself which fix all the points coming from Σ_1. The result of collapsing all such points in Σ_{g+1} together is diffeomorphic to Σ_g, so there is a homomorphism from Γ_{g+1,Σ_1} to Γ_g as well as an inclusion of Γ_{g+1,Σ_1} in Γ_{g+1}. Harer showed that both of these induce isomorphisms

$$H^k(\Gamma_g; \mathbb{Z}) \cong H^k(\Gamma_{g+1,\Sigma_1}; \mathbb{Z}) \text{ and } H^k(\Gamma_{g+1}; \mathbb{Z}) \cong H^k(\Gamma_{g+1,\Sigma_1}; \mathbb{Z})$$

when $g \gg k$, and likewise we have

$$H^k(\Gamma_{g,n}; \mathbb{Z}) \cong H^k(\Gamma_{g+1,\Sigma_1,n}; \mathbb{Z}) \text{ and } H^k(\Gamma_{g+1,n}; \mathbb{Z}) \cong H^k(\Gamma_{g+1,\Sigma_1,n}; \mathbb{Z}) \qquad (3.1)$$

when $g \gg k$.

A similar construction can be made to describe the stabilisation isomorphism

$$H^k(\mathcal{M}_{g,n}; \mathbb{Q}) \cong H^k(\mathcal{M}_{g+1,n}; \mathbb{Q})$$

for the moduli spaces $\mathcal{M}_{g,n}$ (cf. [28, p.31]). Identifying the last marked point of a smooth nonsingular complex projective curve of genus g with a marked point on a curve of genus 1 gives a stable curve of genus $g+1$ with n marked points. This defines for us a morphism

$$\phi : \mathcal{M}_{g,n+1} \times \mathcal{M}_{1,1} \to \overline{\mathcal{M}}_{g+1,n}$$

whose image is an open subset of an irreducible component of the boundary of $\mathcal{M}_{g+1,n}$ in $\overline{\mathcal{M}}_{g+1,n}$, and there is a normal bundle $\mathcal{N}_\phi$ which is a complex line bundle (in the sense of orbifolds) over $\mathcal{M}_{g,n+1} \times \mathcal{M}_{1,1}$. Using $\mathcal{N}_\phi^*$ to denote the complement of the zero section of $\mathcal{N}_\phi$ we can compose projection maps with the forgetful map from $\mathcal{M}_{g,n+1}$ to $\mathcal{M}_{g,n}$ to get

$$\mathcal{N}_\phi^* \to \mathcal{M}_{g,n+1} \times \mathcal{M}_{1,1} \to \mathcal{M}_{g,n+1} \to \mathcal{M}_{g,n}$$

which induces

$$H^k(\mathcal{M}_{g,n}; \mathbb{Q}) \to H^k(\mathcal{N}_\phi^*; \mathbb{Q}). \tag{3.2}$$

On the other hand, using a tubular neighbourhood of the image of ϕ in $\overline{\mathcal{M}}_{g+1,n}$ we obtain a natural homotopy class of maps from $\mathcal{N}_\phi^*$ to $\mathcal{M}_{g+1,n}$ which induces

$$H^k(\mathcal{M}_{g+1,n}; \mathbb{Q}) \to H^k(\mathcal{N}_\phi^*; \mathbb{Q}). \tag{3.3}$$

Here (3.2) and (3.3) represent Harer's maps (3.1) and hence they are isomorphisms if $g \gg k$.

4. Tautological classes

When $g \geq 2$ Mumford [93] and Morita [84] independently defined tautological classes

$$\kappa_i \in H^{2i}(\overline{\mathcal{M}}_g; \mathbb{Q}) \text{ and } e_i \in H^{2i}(\Gamma_g; \mathbb{Z})$$

which correspond up to a sign $(-1)^{i+1}$ in $H^*(\mathcal{M}_g; \mathbb{Q})$ under the isomorphism (2.1). The subalgebra $R^*(\mathcal{M}_g)$ of $H^*(\mathcal{M}_g; \mathbb{Q})$ generated by the κ_i, or equivalently by the e_i, is called its tautological ring.

The classes κ_i are defined using the natural forgetful map $\pi : \mathcal{M}_{g,1} \to \mathcal{M}_g$ which takes an element $[\Sigma, p]$ of $\mathcal{M}_{g,1}$ represented by a nonsingular complex projective curve Σ with one marked point p to the element $[\Sigma]$ of $\mathcal{M}_g$ represented by Σ. This is often called the universal curve over $\mathcal{M}_g$, since for generic choices of Σ the fibre $\pi^{-1}([\Sigma])$ is a copy of Σ. However if Σ has nontrivial automorphisms then $\pi^{-1}([\Sigma])$ is not a copy of Σ but is instead the quotient of Σ by its automorphism group $\mathrm{Aut}(\Sigma)$ (which has size at most $84(g-1)$ when $g \geq 2$).

From the topologists' viewpoint the rôle of $\pi : \mathcal{M}_{g,1} \to \mathcal{M}_g$ is played by the universal oriented Σ_g-bundle

$$\Pi : E\mathrm{Diff}_+\Sigma_g \to B\mathrm{Diff}_+\Sigma_g.$$

Its relative tangent bundle is an oriented real vector bundle of rank 2 on $E\mathrm{Diff}_+\Sigma_g$ (whose fibre at $x \in E\mathrm{Diff}_+\Sigma_g$ is the tangent space at x to the oriented surface $\Pi^{-1}(x)$), so it has an Euler class $e \in H^2(E\mathrm{Diff}_+\Sigma_g; \mathbb{Z})$. Morita defined his tautological classes

$$e_i \in H^{2i}(\Gamma_g; \mathbb{Z}) \cong H^{2i}(B\mathrm{Diff}_+\Sigma_g; \mathbb{Z})$$

by setting e_i to be the pushforward (or integral over the fibres) $\Pi_!(e^{i+1})$ of e^{i+1}.

To define his tautological classes κ_i Mumford used essentially the same procedure with the forgetful map $\pi : \mathcal{M}_{g,1} \to \mathcal{M}_g$, except that he used cotangent spaces instead of tangent spaces (which is the reason that κ_i and e_i only correspond up to a sign $(-1)^{i+1}$) and the relative cotangent bundle (or relative dualising sheaf) for $\pi : \mathcal{M}_{g,1} \to \mathcal{M}_g$ exists as a complex line bundle over $\mathcal{M}_{g,1}$ only in the sense of orbifold line bundles (or line bundles over stacks) because of the existence of nontrivial automorphism groups $\mathrm{Aut}(\Sigma)$.

The forgetful map $\pi : \mathcal{M}_{g,1} \to \mathcal{M}_g$ can be generalised to $\pi : \mathcal{M}_{g,n+1} \to \mathcal{M}_{g,n}$ for any $n \geq 0$ by forgetting the last marked point of an $n+1$-pointed curve, and this can be extended to $\pi : \overline{\mathcal{M}}_{g,n+1} \to \overline{\mathcal{M}}_{g,n}$. Care is needed here when the last marked point lies on an irreducible component with genus 0 and only two other special points; such an irreducible component needs to be collapsed in order to produce a stable n-pointed curve of genus g. This collapsing procedure gives us a forgetful map $\pi : \overline{\mathcal{M}}_{g,n+1} \to \overline{\mathcal{M}}_{g,n}$ whose fibre at $[\Sigma, p_1, \ldots, p_n] \in \overline{\mathcal{M}}_{g,n}$ can be identified with the quotient of Σ by the automorphism group of $(\Sigma, p_1, ..., p_n)$. Mumford's tautological classes can be extended to classes $\kappa_i \in H^{2i}(\overline{\mathcal{M}}_{g,n}; \mathbb{Q})$ (in fact to classes in the rational Chow ring of $\overline{\mathcal{M}}_{g,n}$) defined by

$$\kappa_i = \pi_!(c_1(\omega_{g,n})^{i+1})$$

where $\omega_{g,n}$ is the relative dualising sheaf of $\pi : \overline{\mathcal{M}}_{g,n+1} \to \overline{\mathcal{M}}_{g,n}$ and $c_1(\omega_{g,n}) \in H^2(\overline{\mathcal{M}}_{g,n}; \mathbb{Q})$ is its first Chern class.

When $n > 0$ there are other interesting tautological classes on $\mathcal{M}_{g,n}$ and $\overline{\mathcal{M}}_{g,n}$ exploited by Witten. The forgetful map $\pi : \overline{\mathcal{M}}_{g,n+1} \to \overline{\mathcal{M}}_{g,n}$ has tautological sections $s_j : \overline{\mathcal{M}}_{g,n} \to \overline{\mathcal{M}}_{g,n+1}$ for $1 \leq j \leq n$ such that $s_j([\Sigma, p_1, \ldots, p_n])$ is the element of $\pi^{-1}([\Sigma, p_1, \ldots, p_n]) = \Sigma/\mathrm{Aut}(\Sigma)$ represented by p_j. The Witten classes $\psi_j \in H^2(\overline{\mathcal{M}}_{g,n}; \mathbb{Q})$ for $j = 1, \ldots, n$ can then be defined by

$$\psi_j = c_1(s_j^*(\omega_{g,n})).$$

Roughly speaking, ψ_j is the first Chern class of the (orbifold) line bundle on $\overline{\mathcal{M}}_{g,n}$ whose fibre at $[\Sigma, p_1, \ldots, p_n]$ is the cotangent space $T^*_{p_j}\Sigma$ to Σ at p_j.

The boundary $\Delta = \overline{\mathcal{M}}_{g,n} \setminus \mathcal{M}_{g,n}$ of $\mathcal{M}_{g,n}$ in $\overline{\mathcal{M}}_{g,n}$ is the union of finitely many divisors which meet transversely in $\overline{\mathcal{M}}_{g,n}$. The intersection of any nonempty set of these divisors is the closure of a subset of $\mathcal{M}_{g,n}$ parametrising stable n-pointed curves of some fixed topological type, and is the image of a finite-to-one map to $\overline{\mathcal{M}}_{g,n}$ from a product of moduli spaces $\prod_k \overline{\mathcal{M}}_{g_k,n_k}$ which glues together stable curves of genus g_k with n_k marked points at certain of the marked points. These glueing maps induce pushforward maps on cohomology

$$H^*(\prod_k \overline{\mathcal{M}}_{g_k,n_k}; \mathbb{Q}) \to H^*(\overline{\mathcal{M}}_{g,n}; \mathbb{Q}) \tag{4.1}$$

and the tautological ring $R^*(\overline{\mathcal{M}}_{g,n}; \mathbb{Q})$ is defined inductively to be the subalgebra of $H^*(\overline{\mathcal{M}}_{g,n}; \mathbb{Q})$ generated by the Mumford classes, the Witten classes and the images of the tautological classes in $H^*(\prod_k \overline{\mathcal{M}}_{g_k,n_k}; \mathbb{Q})$ under the pushforward maps (4.1) from the boundary of $\mathcal{M}_{g,n}$. Its restriction to $H^*(\mathcal{M}_{g,n}; \mathbb{Q})$ is the tautological ring of $\mathcal{M}_{g,n}$ and is generated by the Mumford and Witten classes.

5. Mumford's conjecture

Mumford's tautological classes $\kappa_i \in H^{2i}(\mathcal{M}_g;\mathbb{Q})$ are preserved by Harer stabilisation when g is sufficiently large, and so they define elements of the stable cohomology $H^*(\mathcal{M}_\infty;\mathbb{Q})$. Mumford conjectured in [93] that $H^*(\mathcal{M}_\infty;\mathbb{Q})$ is freely generated by $\kappa_1, \kappa_2, \ldots$, or in other words that the obvious map

$$\mathbb{Q}[\kappa_1, \kappa_2, \ldots] \to H^*(\mathcal{M}_\infty;\mathbb{Q}) \tag{5.1}$$

is an isomorphism. Miller [83] and Morita [85] soon proved that this map is injective, so it remained to prove surjectivity. Not long ago Madsen and Tillmann [80] found a homotopy version of Mumford's map (5.1) which they conjectured to be a homotopy equivalence, and very recently Madsen and Weiss [81] have announced a proof of their conjecture, from which Mumford's conjecture follows.

The Madsen-Tillmann map involves the stable mapping class group Γ_∞ rather than the moduli spaces $\mathcal{M}_g$. From the description of $\mathcal{M}_g$ as the quotient of the Γ_g action on Teichmüller space $\mathcal{T}_g$ it follows that when $g \geq 2$ there is a continuous map

$$B\Gamma_g \to \mathcal{M}_g \tag{5.2}$$

uniquely determined up to homotopy. It is known that Γ_g is a perfect group when $g \geq 3$ [46], so we can apply Quillen's plus construction to $B\Gamma_g$ to obtain a simply connected space $B\Gamma_g^+$ with the same homology as $B\Gamma_g$. The moduli space $\mathcal{M}_g$ is also simply connected, so (5.2) factors through a map $B\Gamma_g^+ \to \mathcal{M}_g$ which induces the isomorphism $H^*(\Gamma_g;\mathbb{Q}) \to H^*(\mathcal{M}_g;\mathbb{Q})$ discussed above at (2.1). Moreover Harer stabilisation gives us maps $B\Gamma_g^+ \to B\Gamma_{g+1}^+$ between simply connected spaces which are homology equivalences (and hence also homotopy equivalences) in a range up to some degree which tends to infinity with g. If $B\Gamma_\infty^+$ denotes the homotopy direct limit of these maps as $g \to \infty$, then Mumford's conjecture becomes the statement that

$$H^*(B\Gamma_\infty^+;\mathbb{Q}) \cong \mathbb{Q}[\kappa_1, \kappa_2, \ldots].$$

The conjecture of Madsen and Tillmann [80] describes the homotopy type of $B\Gamma_\infty^+$ (or rather $\mathbb{Z} \times B\Gamma_\infty^+$), giving Mumford's conjecture as a corollary.

Tillmann [103] had already shown that $\mathbb{Z} \times B\Gamma_\infty^+$ is an infinite loop space, in the sense that there exists a sequence of spaces E_n with $E_n = \Omega E_{n+1}$ and $\mathbb{Z} \times B\Gamma_\infty^+ = E_0$. This was an encouraging result because infinite loop spaces have many good properties. Subsequently Madsen and Tillmann [80] found an Ω^∞ map α_∞ from $\mathbb{Z} \times B\Gamma_\infty^+$ to an infinite loop space which they denoted by $\Omega^\infty\mathbb{CP}_{-1}^\infty$ and whose connected component has rational cohomology isomorphic to $\mathbb{Q}[\kappa_1, \kappa_2, \ldots]$.

The infinite loop space $\Omega^\infty\mathbb{CP}_{-1}^\infty$ is related to the limit $\mathbb{CP}^\infty$ of the complex projective spaces $\mathbb{CP}^k$ as $k \to \infty$. Over $\mathbb{CP}^k$ there is a tautological complex line bundle L_k, whose fibre at $x \in \mathbb{CP}^k$ is the one-dimensional subspace of $\mathbb{C}^{k+1}$ represented by x, and a complex vector bundle $L_k^\perp$ of rank k which is its complement in the trivial bundle of rank $k+1$ over $\mathbb{CP}^k$. The restriction of $L_{k+1}^\perp$ to $\mathbb{CP}^k$ is the direct sum of $L_k^\perp$ and a trivial complex line bundle, giving us maps $\mathrm{Th}(L_k^\perp) \to \Omega^2\mathrm{Th}(L_{k+1}^\perp)$ and $\Omega^{2k+2}\mathrm{Th}(L_k^\perp) \to \Omega^{2k+4}\mathrm{Th}(L_{k+1}^\perp)$ where $\mathrm{Th}(L_k^\perp)$ is the Thom space (or one-point

compactification) of the bundle $L_k^\perp$. Madsen and Tillmann define $\Omega^\infty\mathbb{CP}^\infty_{-1}$ to be the direct limit of the spaces $\Omega^{2k+2}\mathrm{Th}(L_k^\perp)$ as $k \to \infty$.

Homotopy classes of maps from an n-dimensional manifold X to $\Omega^\infty\mathbb{CP}^\infty_{-1}$ are represented by proper maps $\phi : M \to X$ from an $(n+2)$-dimensional manifold M together with an 'artificial differential' $\Phi : TM \to \phi^*TX$ and an orientation of $\ker\Phi$. Here Φ is a stable vector bundle surjection; that is, it may be that Φ is defined and becomes a surjective bundle map only once a trivial bundle of sufficiently large rank has been added to TM and ϕ^*TX. Any smooth oriented surface bundle $\phi : E \to X$ induces a homotopy class of maps from X to $\Omega^\infty\mathbb{CP}^\infty_{-1}$ represented by ϕ together with its differential $\Phi = d\phi : TE \to \phi^*TX$, and this effectively defines the Madsen-Tillmann map $\alpha_\infty : \mathbb{Z} \times B\Gamma^+_\infty \to \Omega^\infty\mathbb{CP}^\infty_{-1}$.

Submersion theory suggests a way to tackle the problem of showing that α_∞ is a homotopy equivalence, but compactness of X creates a difficulty for this. Therefore Madsen and Weiss replace X with $X \times \mathbb{R}$. They study a commutative diagram

$$\begin{array}{ccccc} \mathcal{V} & \to & \mathcal{W} & \to & \mathcal{W}_{\mathrm{loc}} \\ \downarrow & & \downarrow & & \downarrow \\ h\mathcal{V} & \to & h\mathcal{W} & \to & h\mathcal{W}_{\mathrm{loc}} \end{array}$$

of contravariant functors from smooth manifolds to sets with the sheaf property for open coverings, and the induced diagram

$$\begin{array}{ccccc} |\mathcal{V}| & \to & |\mathcal{W}| & \to & |\mathcal{W}_{\mathrm{loc}}| \\ \downarrow & & \downarrow & & \downarrow \\ |h\mathcal{V}| & \to & |h\mathcal{W}| & \to & |h\mathcal{W}_{\mathrm{loc}}| \end{array}$$

of the associated spaces, where homotopy classes of maps from X to $|\mathcal{F}|$ correspond naturally to concordance classes in $\mathcal{F}(X)$, and $s_0, s_1 \in \mathcal{F}(X)$ are concordant if $s_0 = t|_{X\times\{0\}}$ and $s_1 = t|_{X\times\{1\}}$ for some $t \in \mathcal{F}(X \times \mathbb{R})$.

If X is any smooth manifold then elements of $\mathcal{V}(X)$ are given by smooth oriented surface bundles E (that is, proper submersions whose fibres are connected oriented surfaces) over $X \times \mathbb{R}$, together with identifications $\partial E \cong \partial(S^1 \times [0,1] \times X \times \mathbb{R})$ compatible with the maps to $X \times \mathbb{R}$. These identifications on the boundary are crucial, because they give $\mathcal{V}$ and the other functors involved the structure of monoids, and thus the associated spaces become topological monoids.

In one version of the bottom row $h\mathcal{V} \to h\mathcal{W} \to h\mathcal{W}_{\mathrm{loc}}$ of the commutative diagram, elements of $h\mathcal{V}(X)$ are given by $(n+3)$-dimensional manifolds E, where $n = \dim X$, and smooth maps $\pi : E \to X$ and $f, g : E \to \mathbb{R}$ such that $(\pi, f) : E \to X \times \mathbb{R}$ is a submersion and $(\pi, g) : E \to X \times \mathbb{R}$ is proper, together with an identification $\partial E \cong \partial(S^1 \times [0,1] \times X \times \mathbb{R})$ compatible with the maps to X and $\mathbb{R}$. If $(\pi, f) : E \to X \times \mathbb{R}$ represents an element of $\mathcal{V}(X)$ then we get an element of $h\mathcal{V}(X)$ by setting $g = f$. The functors $\mathcal{W}$ and $h\mathcal{W}$ are defined similarly, except that the requirement that $(\pi, f) : E \to X \times \mathbb{R}$ should be a submersion is weakened to the requirements that $\pi : E \to X$ should be a submersion and that the restriction of $f : E \to \mathbb{R}$ to any fibre of π should be a Morse function. For $\mathcal{W}_{\mathrm{loc}}$ and $h\mathcal{W}_{\mathrm{loc}}$ the requirements are weakened again, so that 'proper' is replaced by 'proper when restricted to the set of singularities of f on fibres of π'.

The strategy of Madsen and Weiss is to deduce that α_∞ is a homotopy equivalence from the following properties of the commutative diagram above:

(i) the first vertical map represents the Madsen-Tillmann map α_∞;
(ii) the second vertical map is a homotopy equivalence (by a corollary to Vassiliev's h-principle [104]);
(iii) the third vertical map is also a homotopy equivalence (by a much easier argument);
(iv) the bottom row is a homotopy fibre sequence;
(v) the top row becomes a homotopy fibre sequence after group completion (using stratifications of $|\mathcal{W}|$ and $|\mathcal{W}_{\mathrm{loc}}|$ and a subtle application of Harer stabilisation).

6. Faber's conjectures

Although Mumford's conjecture tells us that the tautological classes κ_i generate the stable cohomology ring $H^*(\mathcal{M}_\infty;\mathbb{Q})$, they do not generate $H^*(\mathcal{M}_g;\mathbb{Q})$ for finite g, and in fact $H^*(\mathcal{M}_g;\mathbb{Q})$ has lots of unstable cohomology (at least when g is large enough). This follows from the calculation of Euler characteristics by Harer and Zagier [51] (see also [71]). They show that the orbifold Euler characteristic of $\mathcal{M}_{g,n}$ is

$$(-1)^{n-1}\frac{(2g+n-3)!}{(2g-2)!}\zeta(1-2g)$$

where ζ denotes the Riemann ζ-function, and their work implies that when $g \geq 15$ the Euler characteristic of $\mathcal{M}_g$ is too large in absolute value for $H^*(\mathcal{M}_g;\mathbb{Q})$ to be generated by $\kappa_1, \kappa_2, \ldots$ (cf. also [41, 76]). Nonetheless the tautological ring $R^*(\mathcal{M}_g)$ generated by $\kappa_1, \kappa_2, \ldots$ has many beautiful properties.

Faber [26] has conjectured that $R^*(\mathcal{M}_g)$ has the structure of the algebraic cohomology ring of a nonsingular complex projective variety of dimension $g-2$. More precisely, he conjectured that

(i) $R^k(\mathcal{M}_g)$ is zero when $k > g-2$ and is one-dimensional when $k = g-2$, and the natural pairing $R^k(\mathcal{M}_g) \times R^{g-2-k}(\mathcal{M}_g) \to R^{g-2}(\mathcal{M}_g)$ is perfect. In addition $R^k(\mathcal{M}_g)$ satisfies the Hard Lefschetz property and the Hodge index theorem with respect to the class κ_1.
(ii) The classes $\kappa_1, \ldots, \kappa_{[g/3]}$ generate $R^*(\mathcal{M}_g)$ with no relations in degrees up to and including $[g/3]$.
(iii) Faber also gave an explicit conjecture for a complete set of relations between these generators (in terms of the proportionalities between monomials in $R^{g-2}(\mathcal{M}_g)$).

When $g \leq 15$ Faber [26] has proved all these conjectures concerning $R^*(\mathcal{M}_g)$, and for general g Looijenga [78] and Faber [27] have shown that $R^k(\mathcal{M}_g)$ is zero when $k > g-2$ and is one-dimensional when $k = g-2$. Their proofs apply to both the cohomological version and the Chow ring version of $R^*(\mathcal{M}_g)$. Using topological methods, Morita [90, 91] has recently proved that the classes $\kappa_1, \ldots, \kappa_{[g/3]}$ generate

the cohomological version of $R^*(\mathcal{M}_g)$ (and the rest of (ii) then follows essentially from [50]).

The mapping class group Γ_g acts naturally on $H_1(\Sigma_g;\mathbb{Z})$ in a way which preserves the intersection pairing. This representation gives us an exact sequence of groups

$$1 \to \mathcal{I}_g \to \Gamma_g \to Sp(2g;\mathbb{Z}) \to 1$$

where $\mathcal{I}_g$ denotes the subgroup of Γ_g which acts trivially on $H_1(\Sigma_g;\mathbb{Z})$ and which is called the Torelli group. In [58, 59, 60, 61, 62] Johnson showed that $\mathcal{I}_g$ is finitely generated for $g \geq 3$ (in contrast with the case $g = 2$ [82]), introduced a surjective homomorphism

$$\tau : \mathcal{I}_g \to \wedge^3 H_1(\Sigma_g;\mathbb{Z})/H_1(\Sigma_g;\mathbb{Z})$$

whose kernel is the subgroup of Γ_g generated by all Dehn twists along separating embedded circles, and used τ to determine the abelianisation of $\mathcal{I}_g$. Morita [88] extended the Johnson homomorphism τ to a representation

$$\rho_1 : \Gamma_g \to (\tfrac{1}{2}\wedge^3 H_1(\Sigma_g;\mathbb{Z})/H_1(\Sigma_g;\mathbb{Z})) \rtimes Sp(2g;\mathbb{Z})$$

of the mapping class group Γ_g. Via the cohomology of semi-direct products this induces

$$\rho_1^* : \mathrm{Hom}(\wedge^* U, \mathbb{Q})^{Sp(2g;\mathbb{Z})} \to H^*(\Gamma_g;\mathbb{Q}) \cong H^*(\mathcal{M}_g;\mathbb{Q})$$

where $U = \wedge^3 H_1(\Sigma_g;\mathbb{Q})/H_1(\Sigma_g;\mathbb{Q})$, and the image of ρ_1^* is the tautological ring $R^*(\mathcal{M}_g)$ [63, 79]. By finding suitable relations in $\mathrm{Hom}(\wedge^* U, \mathbb{Q})^{Sp(2g;\mathbb{Z})}$ and exploiting the map $H_1(\Sigma_g;\mathbb{Q}) \to H_1(\Sigma_{g-1};\mathbb{Q})$ induced by collapsing a handle of Σ_g, Morita [90, 91] is able to prove that the classes $\kappa_1, \ldots, \kappa_{[g/3]}$ generate the cohomological version of $R^*(\mathcal{M}_g)$.

Faber, Getzler, Hain, Looijenga, Pandharipande, Vakil and others (cf. [23, 24, 25, 31, 42, 44, 78]) have also made conjectures about the structure of the tautological rings of the compact moduli spaces $\overline{\mathcal{M}}_{g,n}$, which are generated not just by the Mumford classes κ_i but also by the Witten classes ψ_j and the pushforwards of tautological classes from the boundary of $\overline{\mathcal{M}}_{g,n}$. For example, it is expected that $R^*(\overline{\mathcal{M}}_{g,n})$ looks like the algebraic cohomology ring of a nonsingular complex projective variety of dimension $3g-3+n$, while Getzler has conjectured that if $g > 0$ then the monomials of degree g or higher in the Witten classes ψ_j should all come from the boundary of $\overline{\mathcal{M}}_{g,n}$ (a cohomological version of this has been proved by Ionel [54]), and Vakil has made a closely related conjecture that any tautological class in $R^k(\overline{\mathcal{M}}_{g,n})$ with $k \geq g$ should come from classes supported on boundary strata corresponding to stable curves with at least $k-g+1$ components of genus 0.

7. The Virasoro conjecture

The geometry of a nonsingular complex projective variety X can be studied by examining curves in X. Intersection theory on moduli spaces of curves in X, or more precisely moduli spaces of maps from curves to X, leads to Gromov-Witten theory and the quantum cohomology of X, with numerous applications in the last decade to enumerative geometry (cf. [14, 32, 71, 72, 73]).

Let us assume for simplicity that $2g-2+n>0$. For any $\beta \in H_2(X;\mathbb{Z})$ there is a moduli space $\mathcal{M}_{g,n}(X,\beta)$ of n-pointed nonsingular complex projective curves Σ of genus g equipped with maps $f:\Sigma \to X$ satisfying $f_*[\Sigma]=\beta$. This moduli space has a compactification $\overline{\mathcal{M}}_{g,n}(X,\beta)$ which classifies 'stable maps' of type β from n-pointed curves of genus g into X [32]. Here a map $f:\Sigma \to X$ from an n-pointed complex projective curve Σ satisfying $f_*[\Sigma]=\beta$ is called stable if Σ has only nodal singularities and $f:\Sigma \to X$ has only finitely many automorphisms, or equivalently every irreducible component of Σ of genus 0 (respectively genus 1) which is mapped to a single point in X by f contains at least 3 (respectively 1) special points. The forgetful map from $\mathcal{M}_{g,n}(X,\beta)$ to $\mathcal{M}_{g,n}$ which sends $[\Sigma,p_1,\ldots,p_n,\ f:\Sigma \to X]$ to $[\Sigma,p_1,\ldots,p_n]$ extends to a forgetful map $\pi:\overline{\mathcal{M}}_{g,n}(X,\beta)\to\overline{\mathcal{M}}_{g,n}$ which collapses components of Σ with genus 0 and at most two special points.

Of course, when X is itself a single point, $\mathcal{M}_{g,n}(X,\beta)$ and $\overline{\mathcal{M}}_{g,n}(X,\beta)$ are simply the moduli spaces $\mathcal{M}_{g,n}$ and $\overline{\mathcal{M}}_{g,n}$. In general $\overline{\mathcal{M}}_{g,n}(X,\beta)$ has more serious singularities than $\overline{\mathcal{M}}_{g,n}$ and may indeed have many different irreducible components with different dimensions (cf. [66]). Nonetheless, it is a remarkable fact [7, 8, 75] that $\overline{\mathcal{M}}_{g,n}(X,\beta)$ has a 'virtual fundamental class' $[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}$ lying in the expected dimension

$$3g-3+n+(1-g)\dim X+\int_\beta c_1(TX)$$

of $\overline{\mathcal{M}}_{g,n}(X,\beta)$. Gromov-Witten invariants (originally developed mainly in the case $g=0$ when $\overline{\mathcal{M}}_{g,n}(X,\beta)$ is more tractable, but now also studied when $g>0$) are obtained by evaluating cohomology classes on $\overline{\mathcal{M}}_{g,n}(X,\beta)$ against this virtual fundamental class.

The cohomology classes used are of two types. Recall that if $1\le j\le n$ the Witten class $\psi_j \in H^2(\overline{\mathcal{M}}_{g,n};\mathbb{Q})$ is the first Chern class of $s_j^*(\omega_{g,n})$, where s_j is the jth tautological section of the forgetful map from $\overline{\mathcal{M}}_{g,n+1}$ to $\overline{\mathcal{M}}_{g,n}$ and $\omega_{g,n}$ is the relative dualising sheaf of this forgetful map. In a similar way, using the forgetful map from $\overline{\mathcal{M}}_{g,n+1}(X,\beta)$ to $\overline{\mathcal{M}}_{g,n}(X,\beta)$, we can define $\Psi_j \in H^2(\overline{\mathcal{M}}_{g,n}(X,\beta);\mathbb{Q})$ (and Ψ_j is not quite the pullback of ψ_j via the forgetful map $\pi:\overline{\mathcal{M}}_{g,n}(X,\beta)\to\overline{\mathcal{M}}_{g,n}$ because of the collapsing process in the definition of π). We can also pull back cohomology classes on X via the evaluation maps $ev_j:\overline{\mathcal{M}}_{g,n}(X,\beta)\to X$ which send a stable map $f:\Sigma\to X$ to the image $f(p_j)$ of the jth marked point p_j of Σ for $1\le j\le n$.

Gromov-Witten invariants for X are given by integrals

$$\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} ev_1^*(\alpha_1)\ldots ev_n^*(\alpha_n)$$

of classes of the second type $ev_j^*(\alpha_j)$, where $\alpha_1,\ldots,\alpha_n \in H^*(X;\mathbb{Q})$, against the virtual fundamental class of $\overline{\mathcal{M}}_{g,n}(X,\beta)$, while descendent Gromov-Witten invariants are of the form

$$\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \Psi_1^{k_1}\ldots\Psi_n^{k_n} ev_1^*(\alpha_1)\ldots ev_n^*(\alpha_n)$$

for nonnegative integers $k_1, \ldots, k_n$, not all zero. More generally, instead of integrating against $[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}$ to get rational numbers one can consider the image in $H^*(\overline{\mathcal{M}}_{g,n};\mathbb{Q})$ of the product $\Psi_1^{k_1}\ldots\Psi_n^{k_n} ev_1^*(\alpha_1)\ldots ev_n^*(\alpha_n)$ under the virtual push-forward map associated to $\pi : \overline{\mathcal{M}}_{g,n}(X,\beta) \to \overline{\mathcal{M}}_{g,n}$.

When X is a single point, the descendent Gromov-Witten invariants reduce to the integrals

$$\int_{\overline{\mathcal{M}}_{g,n}} \psi_1^{k_1}\ldots\psi_n^{k_n}.$$

Witten [107] conjectured relations between these integrals (later proved by Kontsevich [71] via a combinatorial description of $\overline{\mathcal{M}}_{g,n}$) which enable them to be calculated recursively. Witten's conjecture can be formulated in terms of the formal power series

$$F_g = \sum_{n\geq 0} \frac{1}{n!} \sum_{k_1,\ldots,k_n \geq 0} \int_{\overline{\mathcal{M}}_{g,n}} \psi_1^{k_1}\ldots\psi_n^{k_n} t_{k_1}\ldots t_{k_n}$$

in $\mathbb{Q}[[t_0,t_1,\ldots]]$: it says that $\exp(\sum_{g\geq 0} F_g)$ satisfies a system of differential equations called the Virasoro relations.

Witten's conjecture has been generalised by Eguchi, Hori and Xiong (with an extension by Katz) [14, 22, 34, 37] to provide relations between Gromov-Witten invariants and their descendents for general nonsingular projective varieties X. Their generalisation is called the Virasoro conjecture for X, since it says that a certain formal expression (the 'total Gromov-Witten potential') Z^X in the Gromov-Witten invariants and their descendents satisfies a system of differential equations

$$\mathcal{L}_k Z^X = 0 \text{ for } k \geq -1$$

where the differential operators $\mathcal{L}_k$ satisfy the commutation relations $[\mathcal{L}_k, \mathcal{L}_\ell] = (k-\ell)\mathcal{L}_{k+\ell}$ and hence span a Lie subalgebra of the Virasoro algebra isomorphic to the Lie algebra of polynomial vector fields in one variable (with $\mathcal{L}_k$ corresponding to $-x^{k+1}\mathrm{d}/\mathrm{d}x$). Dubrovin and Zhang [18] have proved that the Virasoro conjecture determines the Gromov-Witten invariants of X when X is homogeneous.

Getzler and Pandharipande [37] showed that part of Faber's conjectures on the structure of the tautological ring of $\mathcal{M}_g$ (the proportionality formulas) would follow from the Virasoro conjecture for $X = \mathbb{CP}^2$, and Givental [39] has recently found a proof of the Virasoro conjecture for a class of varieties which includes all complex projective spaces, thus completing the proof of the proportionality formulas.

Other methods for finding relations between Gromov-Witten invariants include the Toda conjecture [35, 36, 96, 97] and exploitation of intersection theory on $\overline{\mathcal{M}}_{g,n}$ and localisation methods [9, 10, 14, 21, 29, 30, 40, 43, 72], which have been very powerful in enumerative geometry.

8. Moduli spaces of bundles over curves

Another very well studied family of moduli spaces is given by the moduli spaces $\mathcal{B}_\Sigma(r,d)$ of stable holomorphic vector bundles E of rank r and degree d over

a fixed nonsingular complex projective curve Σ of genus $g \geq 2$. When r and d are coprime $\mathcal{B}_\Sigma(r,d)$ is a nonsingular complex projective variety; when r and d have a common factor then $\mathcal{B}_\Sigma(r,d)$ is nonsingular but not projective, and it has a natural compactification $\overline{\mathcal{B}}_\Sigma(r,d)$ which is projective but singular (except when $g = r = 2$) [38, 95]. If the curve Σ is allowed to vary as well as the bundle E over Σ then we obtain a 'universal' moduli space of bundles $\mathcal{B}_g(r,d)$, which maps to the moduli space $\mathcal{M}_g$ of nonsingular curves of genus g with fibre $\mathcal{B}_\Sigma(r,d)$ over $[\Sigma]$. Pandharipande [98] has shown that $\mathcal{B}_g(r,d)$ has a compactification $\overline{\mathcal{B}}_g(r,d)$ which maps to $\overline{\mathcal{M}}_g$ with the fibre over $[\Sigma] \in \mathcal{M}_g$ given by $\overline{\mathcal{B}}_\Sigma(r,d)$.

In the case when r and d are coprime we have a good understanding of the structure of the cohomology ring $H^*(\mathcal{B}_\Sigma(r,d);\mathbb{Z})$, and this understanding is particularly thorough when $r = 2$ [6, 67, 100, 109]. For arbitrary r it is known that the cohomology has no torsion [4] and inductive formulas [4, 16, 45] as well as explicit formulas [5, 74] for computing the Betti numbers are available. There is a simple set of generators for the cohomology ring [4] and there are explicit formulas for the intersection pairings between polynomial expressions in these generators, which in principle determine all the relations by Poincaré duality [17, 56, 101]. There is also an elegant description of a complete set of relations among the generators when $r = 2$ [6, 67, 100, 109], partially motivated by a conjecture of Mumford [69], and there is a generalisation when $r > 2$ which is somewhat less elegant [19].

When r and d are not coprime the structure of the cohomology ring $H^*(\mathcal{B}_\Sigma(r,d);\mathbb{Z})$ is a little more difficult to describe; for example, the induced Torelli group action on $H^*(\mathcal{B}_\Sigma(r,d);\mathbb{Q})$ is nontrivial [13], whereas when r and d are coprime the Torelli action is trivial and the mapping class group acts via representations of $Sp(2g;Z)$ which are easy to determine. However even in this case information is available on the intersection cohomology of the compactification $\overline{\mathcal{B}}_\Sigma(r,d)$ of $\mathcal{B}_\Sigma(r,d)$ and the cohomology of another compactification $\tilde{\mathcal{B}}_\Sigma(r,d)$ of $\mathcal{B}_\Sigma(r,d)$ with only orbifold singularities: for example, there are formulas for the Betti numbers in both cases [68] and their intersection pairings [57, 65], and the mapping class group again acts via representations of $Sp(2g;\mathbb{Z})$ [94].

One of the main reasons for our good understanding of the moduli spaces $\mathcal{B}_\Sigma(r,d)$ (and their compactifications $\overline{\mathcal{B}}_\Sigma(r,d)$ and $\tilde{\mathcal{B}}_\Sigma(r,d)$ when r and d have a common factor) is that they can be constructed as quotients, in the sense of geometric invariant theory [92], of well behaved spaces whose properties are relatively easy to understand. Similar techniques could in principle be used to study the moduli spaces of stable curves $\overline{\mathcal{M}}_g$ and $\overline{\mathcal{M}}_{g,n}$, as well as Pandharipande's compactification $\overline{\mathcal{B}}_g(r,d)$ of the universal moduli space of bundles $\mathcal{B}_g(r,d)$, since they too can be constructed using geometric invariant theory. In practice this has not succeeded except in very special cases because, in contrast to the case of $\mathcal{B}_\Sigma(r,d)$, we do not have quotients of well behaved spaces which are easy to analyse. However as our understanding of the moduli spaces $\overline{\mathcal{M}}_{g,n}(X,\beta)$ of stable maps becomes increasingly well developed, and in particular localisation techniques are used with greater and greater effect, perhaps the techniques available for studying the cohomology of geometric invariant theoretic quotients will provide an additional approach to the cohomology of the moduli spaces $\overline{\mathcal{M}}_g$ and $\overline{\mathcal{M}}_{g,n}$ which can be added to the plethora

of methods already available.

References

[1] E Arbarello and M Cornalba, *Combinatorial and algebro-geometric cohomology classes on the moduli spaces of curves*, J Algebraic Geometry 5 (1996) 705–749.

[2] E Arbarello and M Cornalba, *Calculating cohomology groups of moduli spaces of curves via algebraic geometry*, arXiv:math.AG/9803001.

[3] E Arbarello, M Cornalba, P Griffiths and J Harris, *Geometry of algebraic curves I*, Grundlehren der Mathematischen Wissenschaften 267, Springer-Verlag, 1985.

[4] M F Atiyah and R Bott, *The Yang-Mills equations over Riemann surfaces*, Philos Trans Roy Soc London Ser A **308** (1982) 523–615.

[5] S del Baño, *On the Chow motive of some moduli spaces*, J Reine Angew Math 532 (2001), 105–132.

[6] V Baranovsky *Cohomology ring of the moduli space of stable vector bundles with odd determinant* Izv Russ Acad Nauk **58 n4** (1994) 204–210.

[7] K Behrend, *Gromov-Witten invariants in algebraic geometry*, Invent Math 127 (1997), 601–617.

[8] K Behrend and B Fantechi, *The intrinsic normal cone*, Invent Math 128 (1997), 45–88.

[9] A Bertram, *Some applications of localization to enumerative problems*, Michigan Math J 48 (2000), 65–75.

[10] A Bertram, *Another way to enumerate rational curves with torus actions*, Invent math 142 (2000), 487–512.

[11] J Birman, *Braids, links and mapping class groups*, Annals of Math Studies 82, Princeton, 1975.

[12] C-F Bödigheimer and R Hain, editors, *Mapping class groups and moduli spaces of Riemann surfaces*, Contemporary Math 150, Amer Math Soc 1993.

[13] S Cappell, R Lee and E Miller, *The action of the Torelli group on the homology of representation spaces is nontrivial*, Topology 39 (2000), 851–871.

[14] D Cox and S Katz, *Mirror symmetry and algebraic geometry*, Math Surveys and Monographs 68, Amer Math Soc 1999.

[15] P Deligne and D Mumford, *The irreducibility of the space of curves of given genus*, Publ IHES 36 (1969), 75–110.

[16] U V Desale and S Ramanan, *Poincaré polynomials of the variety of stable bundles* Math. Ann. **216** (1975) 233–244.

[17] S Donaldson *Gluing techniques in the cohomology of moduli spaces* in *Topological methods in modern mathematics* (Proceedings of 1991 Stony Brook conference in honour of the sixtieth birthday of J.Milnor) Publish or Perish.

[18] B Dubrovin and Y Zhang, *Normal forms of hierarchies of integrable PDEs, Frobenius manifolds and Gromov-Witten invariants*, arXiv:math.DG/0108160.

[19] R Earl and F Kirwan, *Complete sets of relations in the cohomology rings of*

moduli spaces of arbitrary rank holomorphic bundles over a Riemann surface, in preparation.

[20] C J Earle and J Eels, *The diffeomorphism group of a compact Riemann surface*, Bull Amer Math Soc 73 (1967), 557–559.

[21] D Edidin and W Graham, *Equivariant intersection theory*, Invent Math 131 (1998), 595–634.

[22] T Eguchi, K Hori and C-S Xiong, *Quantum cohomology and Virasoro algebra* Phys Lett B402 (1997), 71–80.

[23] C Faber, *Chow rings of moduli spaces of curves I: the Chow ring of* $\overline{\mathcal{M}}_3$, Annals of Math 132 (1990), 331–419.

[24] C Faber, *Chow rings of moduli spaces of curves II: some results on the Chow ring of* $\overline{\mathcal{M}}_4$, Annals of Math 132 (1990), 421–449.

[25] C Faber, *Algorithms for computing intersection numbers on moduli spaces of curves, with an application to the class of the locus of Jacobians*, arXiv:alg-geom/9706006v2.

[26] C Faber, *A conjectural description of the tautological ring of the moduli space of curves*, in [28] 109–129.

[27] C Faber, *A non-vanishing result for the tautological ring of* $\mathcal{M}_g$, arXiv:math.AG/9711219.

[28] C Faber and E Looijenga, *Moduli of curves and abelian varieties (the Dutch intercity seminar on moduli)*, Aspects of Mathematics E 33, Vieweg, 1999.

[29] C Faber and R Pandharipande, *Hodge integrals, partition matrices and the* λ_g *conjecture*, arXiv:math.AG/9908052.

[30] C Faber and R Pandharipande, *Hodge integrals and Gromov-Witten theory*, Invent Math 139 (2000), 173–199.

[31] C Faber and R Pandharipande (with an appendix by D Zagier), *Logarithmic series and Hodge integrals in the tautological ring*, arXiv:math.AG/0002112v3.

[32] W Fulton and R Pandharipande, *Notes on stable maps and quantum cohomology*, in *Algebraic Geometry, Santa Cruz 1995*, Proc Symp Pure Math 62 vol 2 (1997), 45–96.

[33] S Gervais, *A finite presentation of the mapping class group of an oriented surface*, arXiv.math.GT/9811162.

[34] E Getzler, *The Virasoro conjecture for Gromov-Witten invariants*, Algebraic Geometry — Hirzebruch 70 (P Pragacz et al, editors), AMS Contemporary Mathematics (1999), arXiv:math.AG/9812026v4.

[35] E Getzler, *The Toda conjecture*, arXiv:math.AG/0108108.

[36] E Getzler, *The equivariant Toda conjecture*, arXiv:math.AG/0207025.

[37] E Getzler and R Pandharipande, *Virasoro constraints and the Chern classes of the Hodge bundle*, Nucl Phys B 530 (1998), 701–714.

[38] D Gieseker, *Geometric Invariant Theory and applications to moduli problems* Proc Int Cong Math (Helsinki, 1978) Academiae Scientarium Fennica (Helsinki, 1980).

[39] A Givental, *Gromov-Witten invariants and quantization of quadratic Hamiltonians*, arXiv:math.AG/0108100.

[40] T Graber and R Pandharipande, *Localization of virtual classes*, Invent Math

135 (1999), 487–518.

[41] T Graber and R Pandharipande, *A non-tautological algebraic class on* $\overline{\mathcal{M}}_{2,22}$, arXiv:math.AG/0104057.

[42] T Graber and R Vakil, *On the tautological ring of* $\overline{\mathcal{M}}_{g,n}$, Proceedings of the 7th Gökova Geometry-Topology conference (2000), arXiv:math. AG/0011100.

[43] T Graber and R Vakil, *Hodge integrals and Hurwitz numbers via virtual localization*, arXiv:math.AG/003028.

[44] R Hain and E Looijenga, *Mapping class groups and moduli space of curves*, in *Algebraic Geometry, Santa Cruz 1995*, Proc Symp Pure Math 62 vol 2 (1997), 97–142.

[45] G Harder and M S Narasimhan, *On the cohomology groups of moduli spaces of vector bundles over curves*, Math Ann 212 (1975), 215–248.

[46] J Harer, *The second homology group of the mapping class group of an orientable surface*, Invent Math 72 (1983), 221–239.

[47] J Harer, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. Math. 121 (1985), 215–251.

[48] J Harer, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent Math 84 (1986), 157–176.

[49] J Harer, *The third homology group of the moduli space of curves*, Duke Math J 63 (1992), 25–55.

[50] J Harer, *Improved stability for the homology of the mapping class groups of surfaces*, Duke University preprint 1993, http://www.math.duke.edu/preprints.

[51] J Harer and D Zagier, *The Euler characteristic of the moduli space of curves*, Invent Math 85 (1986), 457–485.

[52] J Harris and I Morrison, *Moduli of curves*, Graduate Texts in Math 187, Springer, 1998.

[53] R Herrera and S Salamon, *Intersection numbers on moduli spaces and symmetries of a Verlinde formula*, Comm Math Phys 188 (1997) 521–534.

[54] E Ionel, *Topological recursive relations in* $H^{2g}(\mathcal{M}_{g,n})$, arXiv:math.AG/9908060.

[55] N Ivanov, *On the homology stability for Teichmüller modular groups:closed surfaces and twisted coefficients*, in [12] 107–136.

[56] L Jeffrey and F Kirwan, *Intersection theory on moduli spaces of holomorphic bundles of arbitrary rank on a Riemann surface*, Annals of Math 148 (1998), 109–196.

[57] L Jeffrey, Y-H Kiem, F Kirwan, J Woolf, *Intersection pairings on singular moduli spaces of bundles over a Riemann surface*, in preparation.

[58] D Johnson, *An abelian quotient of the mapping class group* $\mathcal{I}_g$, Math Ann 249 (1980), 225–242.

[59] D Johnson, *A survey of the Torelli group*, Contemporary Math 20 (1983), 165–179.

[60] D Johnson, *The structure of the Torelli group I: a finite set of generators for* $\mathcal{I}_g$, Annals of Math 118 (1983), 423–442.

[61] D Johnson, *The structure of the Torelli group II: a characterization of the*

group generated by twists on bounding curves, Topology 24 (1985), 113–126.

[62] D Johnson, *The structure of the Torelli group III: the abelianization of $\mathcal{I}_g$*, Topology 24 (1985), 127–144.

[63] N Kawazumi and S Morita, *The primary approximation to the cohomology of the moduli space of curves and cocycles for the stable characteristic classes*, Math Res Letters 3 (1996), 629–641.

[64] S Keel and S Mori, *Quotients by groupoids*, Annals of Math (2) 145 (1997), 193–213.

[65] Y-H Kiem, *Intersection cohomology of representation spaces of surface groups*, arXiv:math.AG/0101256.

[66] B Kim and R Pandharipande, *The connectedness of the moduli space of maps to homogeneous spaces*, arXiv:math.AG/0003168.

[67] A King and P Newstead, *On the cohomology of the moduli space of rank 2 vector bundles on a curve* Topology **37** (1998) 407–418.

[68] F Kirwan, *On the homology of compactifications of moduli spaces of vector bundles over a Riemann surface*, Proc London Math Soc 53 (1986), 237–266.

[69] F Kirwan, *Cohomology rings of moduli spaces of bundles over Riemann surfaces* J Amer Math Soc **5** (1992) 853–906.

[70] J Kollár, *Quotient spaces modulo algebraic groups*, Annals of Math (2) 145 (1997), 33–79.

[71] M Kontsevich, *Intersection theory on the moduli space of curves and the matrix Airy function*, Commun. Math. Phys. 147 (1992), 1–23.

[72] M Kontsevich, *Enumeration of rational curves via torus actions*, in *The moduli space of curves* (R Dijkgraaf et al, editors), Birkhäuser (1995), 335–368.

[73] M Kontsevich and Y Manin, *Gromov-Witten classes, quantum cohomology and enumerative geometry*, Comm Math Phys 164 (1994), 525–562.

[74] G Laumon and M Rapoport, *The Langlands lemma and the Betti numbers of stacks of G-bundles on a curve*, Int J Math 7 (1996), 29–45.

[75] J Li and G Tian, *Virtual moduli cycles and Gromov-Witten invariants of general symplectic manifolds*, arXiv:alg-geom/9608032v3.

[76] E Looijenga, *Cohomology of $\mathcal{M}_3$ and $\mathcal{M}_3^1$*, in [12] 205–228.

[77] E Looijenga, *Smooth Deligne-Mumford compactifications by means of Prym level structures*, J Algebraic Geometry 3 (1994), 283–293.

[78] E Looijenga, *On the tautological ring of $\mathcal{M}_g$*, Invent Math 121 (1995), 411–419.

[79] E Looijenga, *Stable cohomology of the mapping class group with symplectic coefficients and of the universal Abel-Jacobi map*, J Algebraic Geometry 5 (1996), 135–150.

[80] I Madsen and U Tillmann, *The stable mapping class group and $Q(\mathbb{C}P^\infty_+)$*, Invent. Math. 145 (2001), 509–544.

[81] I Madsen and M Weiss, *Cohomology of the stable mapping class group*, in preparation.

[82] G Mess, *The Torelli groups for genus 2 and 3 surfaces*, Topology 31 (1992), 775–790.

[83] E Miller, *The homology of the mapping class group*, J Diff Geom 24 (1986),

1–14.

[84] S Morita, *Characteristic classes of surface bundles*, Bull Amer Math Soc 11 (1984), 386–388.

[85] S Morita, *Characteristic classes of surface bundles*, Invent Math 90 (1987), 551–577.

[86] S Morita, *Casson's invariant for homology 3-spheres and characteristic classes of surface bundles I*, Topology 28 (1989), 305–323.

[87] S Morita, *On the structure of the Torelli group and Casson's invariant*, Topology 30 (1991), 603–621.

[88] S Morita, *The extension of Johnson's homomorphism from the Torelli group to the mapping class group*, Invent Math 111 (1993), 197–224.

[89] S Morita, *Casson invariant, signature defect of framed manifolds and the secondary characteristic classes of surface bundles*, J Diff Geom 47 (1997), 560–599.

[90] S Morita, *Structure of the mapping class groups of surfaces: a survey and a prospect*, in *Proceedings of the Kirbyfest*, Geometry & Topology Monographs 2 (1999), 349–406.

[91] S Morita, *Generators for the tautological algebra of the moduli space of curves*, to appear in Topology.

[92] D Mumford, *Geometric Invariant Theory*, Ergebnisse der Mathematik 34, Springer, 1965 (3rd enlarged edition with J. Fogarty, F Kirwan, 1994).

[93] D Mumford, *Towards an enumerative geometry of the moduli space of curves*, in *Arithmetic and Geometry Part II*, M Artin and J Tate, editors, Birkhäuser, 1983, 271–328.

[94] G Nelson, *Casson invariants of cobordisms*, Oxford DPhil thesis 1994.

[95] P E Newstead, *Introduction to moduli problems and orbit spaces*, Tata Institute Lecture Notes 51, 1978.

[96] A Okounkov and R Pandharipande, *Gromov-Witten theory, Hurwitz numbers and matrix models I*, arXiv:math.AG/0101147v2.

[97] A Okounkov and R Pandharipande, *Gromov-Witten theory, Hurwitz theory and completed cycles*, arXiv:math.AG/0204305.

[98] R Pandharipande, *A compactification over $\overline{\mathcal{M}}_g$ of the universal moduli space of slope-semistable vector bundles*, J Amer Math Soc 9 (1996), 425–471.

[99] B Riemann, *Theorie der Abel'schen Funktionen*, J. Reine angew. Math. 54 (1857), 115–155.

[100] B Siebert and G Tian, *Recursive relations for the cohomology ring of moduli spaces of stable bundles*, Tr J of Math 19 (1995), 131–144.

[101] M Thaddeus, *Conformal field theory and the cohomology of the moduli space of stable bundles*, J Diff Geom 35 (1992), 131–149.

[102] U Tillmann, *On the homology of the stable mapping class group*, Invent Math 130 (1997), 257–275.

[103] U Tillmann, *Strings and the stable cohomology of mapping class groups*, these proceedings.

[104] V Vassiliev, *Complements of discriminants of smooth maps: topology and applications*, Amer Math Soc Trans Math Monographs 98, 1992.

[105] E Viehweg, *Quasi-projective moduli for polarized varieties*, Ergebnisse der Mathematik 30, Springer, 1995.

[106] B Wajnryb, *A simple presentation for the mapping class group of an orientable surface*, Israel J Math 45 (1983), 157–174.

[107] E Witten, *Two-dimensional gravity and intersection theory on moduli space*, Surveys in Diff. Geom. 1 (1991), 243–310.

[108] E Witten, *Two dimensional gauge theories revisited* J Geom Phys 9 (1992), 303–368.

[109] D Zagier, *On the cohomology of moduli spaces of rank 2 vector bundles over curves*, in *The Moduli space of Curves*, Progress in Mathematics 129, Birkhäuser, 1995, 533–563.

[110] H Zieschang, E Vogt, H-D Coldewey, *Surfaces and planar discontinuous groups*, LNM 835, Springer 1980.

ICM 2002 · Vol. I · 383–400

Chtoucas de Drinfeld, Formule des Traces d'Arthur-Selberg et Correspondance de Langlands

Laurent Lafforgue*

Résumé

Cet exposé est consacré à la preuve de la correspondance de Langlands pour les groupes GL_r sur les corps de fonctions.

Code matière (AMS 2000): 11F, 11F52, 11F60, 11F66, 11F70, 11F72, 11F80, 11R39, 14G35, 14H60, 22E55.
Mots-Clés: Chtoucas, Variétés modulaires de Drinfeld, Modules des fibrés sur les courbes, Correspondance de Langlands, Corps de fonctions, Représentations galoisiennes, Représentations automorphes, Fonctions L, Cohomologie ℓ-adique, Correspondances de Hecke, Formule des traces d'Arthur-Selberg, Formule des points fixes de Grothendieck-Lefschetz.

Introduction

La géométrie algébrique sur un corps F consiste à étudier les variétés algébriques sur F (c'est-à-dire principalement les objets géométriques définis dans des espaces projectifs par des systèmes d'équations polynomiales à coefficients dans F) et les morphismes ou plus généralement les correspondances qui les relient.

Si ℓ est un nombre premier différent de la caractéristique de F, Grothendieck a associé à toute variété algébrique V sur F ses espaces de cohomologie ℓ-adique $H_c^\nu(V \otimes \overline{F}, \mathbb{Q}_\ell)$, $0 \leq \nu \leq 2 \dim V$, qui sont des représentations du groupe de Galois G_F de F continues et de dimension finie sur $\mathbb{Q}_\ell$. Quand le corps F est de type fini sur $\mathbb{Q}$ ou $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, la théorie conjecturale des motifs de Grothendieck et les conjectures de Tate prévoient qu'il devrait être possible de remonter des représentations ℓ-adiques de G_F aux variétés projectives lisses sur F et à leurs correspondances si bien que la connaissance de celles-ci serait essentiellement équivalente à celle des représentations ℓ-adiques irréductibles de G_F.

*Institut des Hautes Etudes Scientifiques, Le Bois Marie, 35 route de Chartres, F-91440 Bures-sur-Yvette, France. Courriel : laurent@ihes.fr

Cette perspective pose de façon cruciale la question de dresser la liste de toutes les représentations ℓ-adiques irréductibles de G_F. Quand F est un "corps global", c'est-à-dire une extension finie de $\mathbb{Q}$ ou bien le corps $F(X)$ des fonctions rationnelles d'une courbe X projective lisse sur un corps fini, Langlands a conjecturé que les représentations irréductibles de dimension r de G_F sont paramétrées naturellement par les représentations automorphes cuspidales de GL_r sur F.

L'exposé est consacré à la preuve de la correspondance de Langlands dans le cas des corps de fonctions $F = F(X)$. Un rôle central y est joué par la formule des traces d'Arthur-Selberg que la théorie des chtoucas de Drinfeld permet d'interpréter géométriquement.

1. L'énoncé de la correspondance de Langlands

On considère donc une courbe X projective, lisse et géométriquement connexe sur un corps $\mathbb{F}_q$ à q éléments. On note $F = F(X)$ le corps des fonctions rationnelles sur X et $|X|$ l'ensemble des points fermés de X. Pour $x \in |X|$, on note κ_x son corps résiduel et $\deg(x)$ la dimension de κ_x sur $\mathbb{F}_q$.

1.1. Représentations galoisiennes ℓ-adiques

On note G_F le groupe de Galois de F, c'est-à-dire le groupe des automorphismes d'une clôture séparable $\overline{F}$ de F. C'est un groupe profini.

Soit ℓ un nombre premier différent de la caractéristique de $\mathbb{F}_q$.

Une représentation ℓ-adique σ de G_F est une représentation continue et de dimension finie sur $\overline{\mathbb{Q}}_\ell$ qui est définie sur une extension finie de $\mathbb{Q}_\ell$ et provient d'un faisceau ℓ-adique lisse (ou "système local") sur un ouvert non vide de X. On dit que σ est non ramifiée en un point $x \in |X|$ si x est dans le plus grand ouvert où σ est un système local. Dans ce cas, la fibre σ_x de σ en x est une représentation du groupe de Galois de κ_x lequel est engendré par l'automorphisme Frob_x d'élévation à la puissance $q^{\deg(x)}$ et on peut poser

$$\mathrm{L}_x(\sigma, T) = \det_{\sigma_x}(1 - T \cdot \mathrm{Frob}_x^{-1})^{-1}\,.$$

Pour tout entier $r \geq 1$, on note $\{\sigma\}_r$ l'ensemble des classes d'isomorphie de représentations ℓ-adiques irréductibles de dimension r de G_F dont le déterminant est d'ordre fini.

1.2. Groupes adéliques et algèbres de Hecke

Tout point $x \in |X|$ induit sur F la valuation

$$\deg_x : F^\times \mapsto \mathbb{Z} \qquad f \mapsto \deg_x(f) = \text{ l'ordre d'annulation de } f \text{ en } x.$$

On note F_x le corps complété de F en x ; il contient comme sous-anneau d'entiers $O_x = \{f_x \in F_x \mid \deg_x(f_x) \geq 0\}$ qui est compact.

L'anneau $\mathbb{A}$ des adèles de F est le produit "restreint" $\prod_{x \in |X|} F_x$ des familles $f_x \in F_x$, $x \in |X|$, telles que $f_x \in O_x$ pour presque tout x (c'est-à-dire tout x sauf

un nombre fini). Il contient comme sous-anneau compact des entiers $O_{\mathbb{A}} = \prod_{x \in |X|} O_x = \varprojlim_N \mathcal{O}_N$, où N décrit l'ensemble des "niveaux" c'est-à-dire des sous-schémas fermés finis $N = \operatorname{Spec} \mathcal{O}_N \hookrightarrow X$.

L'anneau $\mathbb{A}$ contient F comme sous-groupe additif discret cocompact et le noyau $\mathbb{A}^{\times 0}$ de l'homomorphisme

$$\deg : \mathbb{A}^\times \to \mathbb{Z} \qquad (f_x) \mapsto \sum_{x \in |X|} \deg(x)\, \deg_x(f_x)$$

contient $F^\times$ comme sous-groupe discret cocompact.

Pour tout entier $r \geq 1$, le groupe topologique $\mathrm{GL}_r(F_x)$ [resp. $\mathrm{GL}_r(\mathbb{A}) = \prod_{x \in |X|} \mathrm{GL}_r(F_x)$] a une mesure de Haar $\mathrm{d}g_x$ [resp. $\mathrm{d}g_{\mathbb{A}} = \prod \mathrm{d}g_x$] qui attribue le volume 1 au sous-groupe ouvert compact maximal $K_x = \mathrm{GL}_r(O_x)$ [resp. $K = \mathrm{GL}_r(O_{\mathbb{A}}) = \prod K_x$]. L'algèbre de Hecke est l'algèbre de convolution $\mathcal{H}^r_x$ [resp. $\mathcal{H}^r = \otimes \mathcal{H}^r_x$] des fonctions localement constantes à support compact sur $\mathrm{GL}_r(F_x)$ [resp. $\mathrm{GL}_r(\mathbb{A})$]. Elle est réunion filtrante des sous-algèbres $\mathcal{H}^r_{N,x}$ [resp. $\mathcal{H}^r_N = \otimes \mathcal{H}^r_{N,x}$] des fonctions bi-invariantes par les sous-groupes ouverts d'indice fini $K_{N,x} = \operatorname{Ker}(K_x \to \mathrm{GL}_r(\mathcal{O}_N))$ [resp. $K_N = \prod K_{N,x} = \operatorname{Ker}(K \twoheadrightarrow \mathrm{GL}_r(\mathcal{O}_N)$]. Chaque $\mathcal{H}^r_{N,x}$ [resp. $\mathcal{H}^r_N$] a une unité $\mathbb{1}_{N,x}$ [resp. $\mathbb{1}_N = \otimes \mathbb{1}_{N,x}$].

On s'intéresse aux représentations "lisses admissibles irréductibles" de $\mathrm{GL}_r(F_x)$ [resp. $\mathrm{GL}_r(\mathbb{A})$]. Ce sont les modules simples π_x sur $\mathcal{H}^r_x$ [resp. π sur $\mathcal{H}^r$] qui sont réunions filtrantes des $\pi_x \cdot \mathbb{1}_{N,x}$ [resp. $\pi \cdot \mathbb{1}_N$] supposés de dimension finie. Chaque $\pi_x \cdot \mathbb{1}_{N,x}$ [resp. $\pi \cdot \mathbb{1}_N$] est un module sur $\mathcal{H}^r_{N,x}$ [resp. $\mathcal{H}^r_N$], et s'il n'est pas nul il est simple et caractérise π_x [resp. π]. L'unité $\mathbb{1}_{\emptyset,x}$ de l'algèbre de Hecke "sphérique" $\mathcal{H}^r_{\emptyset,x}$ est la fonction caractéristique de $K_x = \mathrm{GL}_r(O_x)$; quand $\pi_x \cdot \mathbb{1}_{\emptyset,x} \neq 0$, on dit que π_x est "non ramifiée". Enfin, les représentations lisses admissibles irréductibles de $\mathrm{GL}_r(\mathbb{A})$ sont celles de la forme $\pi = \bigotimes_{x \in |X|} \pi_x$ où les π_x sont des représentations lisses admissibles irréductibles des $\mathrm{GL}_r(F_x)$ presque toutes non ramifiées.

Théorème (Satake). – *L'algèbre sphérique $\mathcal{H}^r_{\emptyset,x}$ est commutative et isomorphe à l'algèbre des polynômes symétriques en $Z_1, Z_1^{-1}, \ldots, Z_r, Z_r^{-1}$.*

Par conséquent, une représentation lisse admissible non ramifiées irréductible π_x de $\mathrm{GL}_r(F_x)$ est caractérisée par r scalaires non nuls $z_1(\pi_x), \ldots, z_r(\pi_x)$ (appelés valeurs propres de Hecke) bien définis à l'ordre près ou par

$$\mathrm{L}(\pi_x, T) = \prod_{i=1}^{r} (1 - T \cdot z_i(\pi_x)^{-1})^{-1}\,.$$

Et si $\pi = \otimes \pi_x$ est une représentation lisse admissible irréductible de $\mathrm{GL}_r(\mathbb{A})$, $\mathrm{L}_x(\pi, T) = \mathrm{L}(\pi_x, T)$ est défini déjà en tout x où π_x est non ramifiée, donc en presque tout x.

1.3. Représentations automorphes cuspidales et décomposition spectrale de Langlands

Pour tout entier $r \geq 1$, $\mathrm{GL}_r(F)$ est un sous-groupe discret de covolume fini (mais non cocompact si $r \geq 2$) dans $\mathrm{GL}_r(\mathbb{A})^0 = \mathrm{Ker}(\mathrm{GL}_r(\mathbb{A}) \xrightarrow{\det} \mathbb{A}^\times \xrightarrow{\deg} \mathbb{Z})$. La théorie automorphe consiste à étudier le quotient

$$\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A})$$

via l'espace de ses fonctions muni de l'action de $\mathcal{H}^r$ par convolution à droite.

Considérons d'abord l'espace des fonctions φ sur $\mathrm{GL}_r(\mathbb{A})$ qui sont localement constantes, à support compact, invariantes à gauche par $\mathrm{GL}_r(F)$ et vérifient $\varphi(ag) = \varphi(g)$, $\forall\, g$, pour un certain $a \in \mathbb{A}^\times$ de degré $\neq 0$ et la condition de "cuspidalité"

$$\int_{N_P(F)\backslash N_P(\mathbb{A})} \varphi(n_P\, g) \cdot \mathrm{d}n_P = 0\,, \quad \forall\, g \in \mathrm{GL}_r(\mathbb{A})\,, \quad \forall\, P \subsetneq \mathrm{GL}_r\,,$$

où P décrit l'ensemble des sous-groupes paraboliques standard de GL_r, N_P désigne le radical unipotent de P et $\mathrm{d}n_P$ une mesure de Haar sur $N_P(\mathbb{A})$.

Il s'écrit comme une somme directe de représentations lisses admissibles irréductibles de $\mathrm{GL}_r(\mathbb{A})$ appelées les représentations automorphes cuspidales. On note $\{\pi\}_r$ leur ensemble. Chaque $\pi \in \{\pi\}_r$ a un caractère central χ_π d'ordre fini.

On doit aussi introduire les paires (Q, π) constituées d'un sous-groupe parabolique standard $Q \subsetneq \mathrm{GL}_r$ associé à une partition $r = r_1 + \cdots + r_k$, avec pour sous-groupe de Lévi $M_Q = \mathrm{GL}_{r_1} \times \cdots \times \mathrm{GL}_{r_k}$, et d'une représentation irréductible π de $M_Q(\mathbb{A})$ qui est le produit $\pi_1 \otimes \ldots \otimes \pi_k$ de représentations automorphes cuspidales $\pi_1, \ldots, \pi_k$ de $\mathrm{GL}_{r_1}(\mathbb{A}), \ldots, \mathrm{GL}_{r_k}(\mathbb{A})$.

Deux paires (Q, π) et (Q', π') sont dites équivalentes s'il existe une permutation σ échangeant les facteurs de $M_Q = \mathrm{GL}_{r_1} \times \ldots \times \mathrm{GL}_{r_k}$ et de $M_{Q'}$ et un caractère λ de $M_Q(\mathbb{A})$ "non ramifié" c'est-à-dire se factorisant à travers l'homomorphisme

$$\mathrm{GL}_{r_1}(\mathbb{A}) \times \ldots \times \mathrm{GL}_{r_k}(\mathbb{A}) \xrightarrow{\det} (\mathbb{A}^\times)^k \xrightarrow{\deg} \mathbb{Z}^k$$

tels que $\pi' \cong \sigma(\lambda \otimes \pi)$. On choisit un représentant (Q, π) dans chaque classe d'équivalence.

Soit $a \in \mathbb{A}^\times$ un adèle inversible de degré non nul. L'espace de Hilbert $L^2(\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}})$ des fonctions de carré intégrable sur $\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}}$ est muni d'une action de l'algèbre de Hecke $\mathcal{H}^r$ par convolution à droite. Le théorème fondamental de la théorie des fonctions automorphes décrit sa décomposition comme somme de représentations irréductibles :

Théorème de décomposition spectrale de Langlands. – *On a*

$$L^2(\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}}) = \bigoplus_{\substack{\pi \in \{\pi\}_r \\ \chi_\pi(a)=1}} \pi \oplus \bigoplus_{(Q,\pi)} [\textit{somme continue de représentations irréductibles construites à partir de } (Q,\pi) \textit{ via la théorie des séries d'Eisenstein}].$$

Nous énonçons ici ce théorème en termes très vagues. En effet, on a seulement besoin de savoir pour la suite que les représentations automorphes cuspidales de GL_r apparaissent dans la somme et que tout le reste provient des rangs $< r$.

1.4. La correspondance de Langlands sur les corps de fonctions

On choisit un isomorphisme algébrique entre $\overline{\mathbb{Q}}_\ell$ et $\mathbb{C}$.

Avec Langlands, disons qu'une représentation automorphe cuspidale π de GL_r et une représentation ℓ-adique σ de G_F irréductible de dimension r se correspondent si, en tout $x \in |X|$ où π est non ramifiée, σ est non ramifiée et

$$\mathrm{L}_x(\sigma, T) = \mathrm{L}_x(\pi, T)\,.$$

Théorème. – *Pour tout entier $r \geq 1$, cette correspondance définit une bijection*

$$\{\pi\}_r \cong \{\sigma\}_r \qquad \pi \mapsto \sigma_\pi\ , \qquad \sigma \mapsto \pi_\sigma\,.$$

Le cas $r = 1$ est la théorie du corps de classes global pour $F = F(X)$.

Le cas $r = 2$ a été démontré par Drinfeld au début des années 70 en inventant les chtoucas et étudiant ceux de rang 2.

Le cas $r \geq 3$ a été démontré par l'auteur il y a deux ans en étudiant les chtoucas de Drinfeld de rang r.

On sait que l'isomorphisme de Satake s'étend en une "correspondance de Langlands locale" qui, pour tous les corps F_x localisés de $F = F(X)$ en les points $x \in |X|$, a été démontrée il y a dix ans par Laumon, Rapoport et Stuhler en étudiant la cohomologie ℓ-adique des variétés modulaires de "$\mathcal{D}$-faisceaux elliptiques". Elle fait se correspondre les représentations ℓ-adiques de dimension r du groupe de Galois G_{F_x} de F_x et les représentations lisses admissibles irréductibles de $\mathrm{GL}_r(F_x)$.

La correspondance globale du théorème ci-dessus est compatible avec la correspondance locale au sens que si $\pi \in \{\pi\}_r$ et $\sigma \in \{\sigma\}_r$ se correspondent au sens global, alors en tout point $x \in |X|$ (y compris ceux où il y a ramification), le facteur local π_x de π en x et la restriction σ_x de σ à G_{F_x} se correspondent au sens local.

2. Formule des traces d'Arthur-Selberg et chtoucas de Drinfeld

On va commencer à expliquer la démonstration de la correspondance de Langlands en rang r. La première chose importante à dire est qu'elle se fait par récurrence. On suppose $r \geq 2$ et la correspondance déjà connue en tous les rangs $r' < r$.

2.1. Des représentations galoisiennes vers les représentations automorphes

Pour $\sigma \in \{\sigma\}_r$, l'unicité de $\pi_\sigma \in \{\pi\}_r$ lui correspondant au sens de Langlands résulte du "théorème de multiplicité 1 fort" de Piatetski-Shapiro qui dit qu'une représentation automorphe cuspidale de GL_r est caractérisée par la connaissance de ses valeurs propres de Hecke en presque tout $x \in |X|$.

Quant à l'existence de π_σ, il est connu depuis les années 80 (avec la "formule du produit" de Laumon) qu'elle résulte de la correspondance de Langlands en les rangs $r' < r$. Rappelons comment :

La théorie des modèles de Whittaker permet d'associer à σ une représentation lisse admissible irréductible π de GL_r qui, en tout x où σ est non ramifiée, est non ramifiée et vérifie $\mathrm{L}_x(\pi, T) = \mathrm{L}_x(\sigma, T)$, et qui est réalisée dans un espace de fonctions φ sur $P_1(F)\backslash \mathrm{GL}_r(\mathbb{A})$, où $P_1 \subsetneq \mathrm{GL}_r$ désigne le sous-groupe parabolique standard de type $r = (r-1)+1$. Le problème est de montrer que toutes ces fonctions φ sont invariantes à gauche par $\mathrm{GL}_r(F)$ tout entier.

D'après les "théorèmes réciproques" de Hecke, Weil (pour GL_2) et Piatetski-Shapiro (pour GL_r, $r \geq 3$), c'est équivalent à montrer que pour tout $r' < r$ et toute représentation automorphe cuspidale $\pi' \in \{\pi\}_{r'}$, la fonction L globale

$$\mathrm{L}(\sigma \times \pi', T)$$

est un polynôme qui vérifie une certaine équation fonctionnelle.

Or, d'après l'hypothèse de récurrence, π' correspond à une représentation galoisienne $\sigma' \in \{\sigma\}_{r'}$ et on peut écrire

$$\mathrm{L}(\sigma \times \pi', T) = \mathrm{L}(\sigma \otimes \sigma', T)\,.$$

Comme on est sur un corps de fonctions $F = F(X)$, on sait grâce à Grothendieck qu'une telle fonction L galoisienne $\mathrm{L}(\sigma \otimes \sigma', T)$ est un polynôme et vérifie une équation fonctionnelle induite par la dualité de Poincaré sur la courbe X. Il faut encore vérifier que la constante dans l'équation fonctionnelle est celle qu'on veut et ceci est la formule du produit de Laumon.

2.2. Quotients adéliques et fibrés vectoriels sur la courbe

Les différentes approches géométriques de Drinfeld pour le programme de Langlands sur les corps de fonctions sont fondées sur la remarque suivante d'André Weil :

Lemme. – *Le quotient double* $\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/K$ *s'identifie à l'ensemble des classes d'isomorphie de fibrés vectoriels de rang* r *sur la courbe* X.

Plus généralement, si $K_N \subseteq K = \mathrm{GL}_r(O_{\mathbb{A}})$ *est le sous-groupe ouvert associé à un niveau* $N \hookrightarrow X$, $\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/K_N$ *s'identifie à l'ensemble des classes de fibrés* $\mathcal{E}$ *de rang* r *sur* X *munis d'une structure de niveau* N *c'est-à-dire d'un isomorphisme* $\mathcal{E} \otimes_{\mathcal{O}_X} \mathcal{O}_N = \mathcal{E}_N \xrightarrow{\sim} \mathcal{O}_X^r$.

Autrement dit, $\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/K$ est l'ensemble des points à valeurs dans $\mathbb{F}_q$ du champ $\mathcal{V}ec^r$ des fibrés vectoriels de rang r sur X.

Cette identification a permis à Drinfeld de construire l'application $\sigma \mapsto \pi_\sigma$ en rang $r = 2$ d'une façon purement géométrique, construction qui, dans le cas non ramifié ($N = \emptyset$), a été progressivement généralisée en rang r arbitraire par Laumon, Kazhdan, Frenkel, Gaitsgory, Vilonen.

On part d'une représentation ℓ-adique partout non ramifiée $\sigma \in \{\sigma\}_r$ qu'on voit comme un système local sur X supposé absolument irréductible. En réinterprétant géométriquement la construction des modèles de Whittaker, on lui associe un certain complexe Aut'_σ de faisceaux ℓ-adiques sur le champ $\mathcal{V}ec'^r$ des fibrés $\mathcal{E}$ de rang r sur X munis d'un plongement $\Omega^1_X \hookrightarrow \mathcal{E}$ du fibré inversible canonique de X. Le quotient $P_1(F)\backslash \mathrm{GL}_r(\mathbb{A})/K$ s'identifie à l'ensemble des points à valeurs dans $\mathbb{F}_q$ d'un certain ouvert de $\mathcal{V}ec'^r$ et la fonction φ qui associe à tout tel point la somme alternée des traces de l'élément de Frobenius agissant sur la fibre de Aut'_σ en ce point est celle associée à σ par la théorie classique des modèles de Whittaker. Le problème est de montrer que Aut'_σ "se descend" par le morphisme $\mathcal{V}ec'^r \to \mathcal{V}ec^r$ d'oubli des plongements de Ω^1_X c'est-à-dire qu'il est l'image réciproque d'un certain faisceau pervers Aut_σ sur $\mathcal{V}ec^r$. Or il y a un grand ouvert de $\mathcal{V}ec^r$ au-dessus duquel le quotient de $\mathcal{V}ec'^r$ par $\mathbb{G}_m$ est un fibré projectif. Comme les espaces projectifs sont simplement connexes, la descente se fait automatiquement si l'on sait que la restriction de Aut'_σ au-dessus de cet ouvert est un système local ($\mathbb{G}_m$-équivariant). Frenkel, Gaitsgory et Vilonen ont montré que cette propriété résulte de l'annulation de certains foncteurs cohomologiques dits "de moyennisation" et Gaitsgory a récemment démontré cet énoncé d'annulation par des arguments purement géométriques.

On voit que cette démonstration de l'existence de l'application $\sigma \mapsto \pi_\sigma$ est profondément différente de celle du paragraphe précédent car ici la descente repose sur la simple connexité des espaces projectifs et non plus sur les propriétés des fonctions L déduites de l'existence des applications $\pi' \mapsto \sigma_{\pi'}$ en rangs $< r$.

2.3. La formule des traces d'Arthur-Selberg

On veut aborder maintenant la construction de l'application $\pi \mapsto \sigma_\pi$. L'unicité des σ_π correspondant aux $\pi \in \{\pi\}_r$ résulte du théorème de densité de Chebotarev.

Pour l'existence, il faut commencer par avoir une prise sur l'ensemble $\{\pi\}_r$. Une telle prise est fournie par la formule des traces d'Arthur-Selberg que nous rappelons :

L'image $\varphi * h$ d'une fonction $\varphi \in L^2(\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}})$ par convolution par

$h \in \mathcal{H}^r$ est donnée par

$$(\varphi * h)(g') = \int_{\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}}} K_{h,G}(g', g)\, \varphi(g) \cdot \mathrm{d}g$$

où

$$K_{h,G}(g', g) = \sum_{\substack{\gamma \in \mathrm{GL}_r(F) \\ n \in \mathbb{Z}}} h(g^{-1} \gamma\, a^n\, g')\,.$$

Pour $r \geq 2$, le quotient $\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}}$ a certes un volume fini mais il n'est pas compact si bien que l'action de $h \in \mathcal{H}^r$ n'a pas de trace au sens que l'intégrale

$$\text{“}\,\mathrm{Tr}(h)\text{”} = \int_{\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}}} K_{h,G}(g, g) \cdot \mathrm{d}g$$

diverge en général. Il résulte du théorème de décomposition spectrale de Langlands que le noyau $K_{h,G}$ s'écrit naturellement

$$K_{h,G} = \sum_{\substack{\pi \in \{\pi\}_r \\ \chi_\pi(a) = 1}} K^{\pi}_{h,G} + \sum_{(Q,\pi)} K^{Q,\pi}_{h,G}$$

et il faut préciser que le problème de divergence provient des $Q \subsetneq G = \mathrm{GL}_r$.

Pour cette raison, on introduit les troncatures d'Arthur :

Pour tout sous-groupe parabolique standard $P \subsetneq G$, l'action de h sur $L^2(P(F)\backslash \mathrm{GL}_r(\mathbb{A})/a^{\mathbb{Z}})$ a un noyau

$$(g', g) \mapsto K_{h,P}(g', g) = \sum_{\substack{\gamma \in P(F) \\ n \in \mathbb{Z}}} h(g^{-1} \gamma\, a^n\, g')$$

qui se décompose naturellement en une somme

$$K_{h,P} = \sum_{(Q,\pi)} K^{Q,\pi}_{h,P}$$

où n'apparaissent que les Q tels que $M_Q \subseteq M_P$ à permutation près des facteurs.

On considère un polygone de troncature, c'est-à-dire une fonction convexe $p : [0, r] \to \mathbb{R}_+$, s'annulant en 0 et r et affine sur chaque intervalle $[r' - 1, r']$, $0 < r' \leq r$. On pose

$$K^{\leq p}_{h,G}(g, g) = K_{h,G}(g, g) + \sum_{P \subsetneq \mathrm{GL}_r = G} (-1)^{|P|-1} \sum_{\delta \in P(F)\backslash G(F)} \mathbb{1}^p_P(\delta g)\, K_{h,P}(\delta g, \delta g)$$

où $|P|$ désigne le nombre de facteurs de chaque M_P et $\mathbb{1}^p_P$ est une fonction caractéristique (prenant les valeurs 1 ou 0) sur $P(F)\backslash G(\mathbb{A})$ qui dépend du polygone de troncature p. Puis on pose

$$\mathrm{Tr}^{\leq p}(h) = \int_{G(F)\backslash G(\mathbb{A})/a^{\mathbb{Z}}} K^{\leq p}_{h,G}(g, g) \cdot \mathrm{d}g\,.$$

On définit aussi des $\mathrm{Tr}^{\leq p}_{Q,\pi}(h)$ par des sommes alternées et intégrales semblables à partir des $K^{Q,\pi}_{h,P}(\cdot,\cdot)$.

Comme conséquence du théorème de décomposition spectrale de Langlands, on montre :

Théorème (formule des traces d'Arthur-Selberg). – *Toutes ces intégrales convergent et on a*

$$\mathrm{Tr}^{\leq p}(h) = \sum_{\substack{\pi \in \{\pi\}_r \\ \chi_\pi(a)=1}} \mathrm{Tr}_\pi(h) + \sum_{(Q,\pi)} \mathrm{Tr}^{\leq p}_{Q,\pi}(h)\,.$$

On voit donc que dans $\mathrm{Tr}^{\leq p}(h)$ apparaissent toutes les traces $\mathrm{Tr}_\pi(h)$ plus d'autres termes (qui peuvent être explicités). Ceux-ci sont très compliqués, ils dépendent du polygone de troncature p et ils ne sont pas invariants par conjugaison de h mais le plus important est qu'ils proviennent des rangs $< r$.

Un autre point important est que les troncatures d'Arthur trouvent un sens géométrique dans l'équivalence de Weil :

Proposition. – *Si le polygone de troncature $p : [0,r] \to \mathbb{R}_+$ est "assez convexe" (c'est-à-dire si les différences de pentes $[p(r') - p(r'-1)] - [p(r'+1) - p(r')]$, $0 < r' < r$, sont assez grandes) en fonction de h, alors pour tout $g \in G(F)\backslash G(\mathbb{A})$ auquel est associé un fibré $\mathcal{E}$ de rang r sur X, on a*

$$K^{\leq p}_{h,G}(g,g) = \begin{cases} K_{h,G}(g,g) & \text{si le polygone canonique de Harder-Narasimhan} \\ & \text{de } \mathcal{E} \text{ est} \leq p, \\ 0 & \text{sinon.} \end{cases}$$

2.4. Les chtoucas de Drinfeld

Pour être exploitée, la formule des traces d'Arthur-Selberg a besoin d'être combinée avec autre chose. Les chtoucas de Drinfeld vont permettre de lui donner une interprétation géométrique et de là une interprétation cohomologique via le théorème des points fixes de Grothendieck-Lefschetz.

Le problème est de construire des représentations ℓ-adiques σ_π de G_F. En géométrie algébrique, on peut construire des représentations galoisiennes en définissant des variétés sur F (ou sur X ou, comme il va se produire, sur $X \times X$) et en prenant leur cohomologie ℓ-adique (ou celle de leur fibre générique). Dans notre situation, il faut bien sûr que ces variétés aient un rapport étroit avec les représentations automorphes et donc avec le quotient $\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})$.

Revenant à l'équivalence de Weil, on remarque que se donner un fibré sur X équivaut à se donner un fibré $\mathcal{E}$ sur $\overline{X} = X \otimes_{\mathbb{F}_q} \overline{\mathbb{F}}_q$ muni d'un isomorphisme avec son transformé ${}^\tau\mathcal{E} = (\mathrm{Id}_X \otimes \mathrm{Frob})^*\mathcal{E}$ par l'endomorphisme de Frobenius. Drinfeld a découvert qu'en autorisant cet isomorphisme $\mathcal{E} \cong {}^\tau\mathcal{E}$ à avoir un pôle et un zéro, on définit un problème de modules dont le classifiant répond à la question posée :

Définition (Drinfeld). – (i) *Un chtouca de rang r à valeurs dans un schéma S (sur le corps de base $\mathbb{F}_q$) consiste en*

- *un fibré $\mathcal{E}$ localement libre de rang r sur $X \times S$,*
- *une modification de $\mathcal{E}$ c'est-à-dire un diagramme*

$$\mathcal{E} \overset{j}{\hookrightarrow} \mathcal{E}' \overset{t}{\hookleftarrow} \mathcal{E}''$$

où $\mathcal{E}', \mathcal{E}''$ sont des fibrés sur $X \times S$ et où j, t sont des plongements dont les conoyaux sont supportés par les graphes de deux morphismes $\infty, 0 : S \to X$ et sont inversibles comme faisceaux cohérents sur $\mathcal{O}_S$,
- *un isomorphisme* $(\mathrm{Id}_X \times \mathrm{Frob}_S)^* \mathcal{E} = {}^{\tau}\mathcal{E} \xrightarrow{\sim} \mathcal{E}''$.

(ii) *Une structure de niveau $N \hookrightarrow X$ sur un tel chtouca (dont le pôle ∞ et le zéro 0 évitent N) est un isomorphisme $\mathcal{E} \otimes_{\mathcal{O}_{X \times S}} \mathcal{O}_{N \times S} = \mathcal{E}_N \xrightarrow{\sim} \mathcal{O}^r_{N \times S}$ compatible avec l'isomorphisme ${}^{\tau}\mathcal{E}_N \xrightarrow{\sim} \mathcal{E}_N$.*

Quand S varie, les groupoïdes de chtoucas de rang r [resp. et avec structures de niveau N] constituent un champ Cht^r [resp. Cht^r_N] dont on montre qu'il est algébrique au sens de Deligne-Mumford et localement de type fini. Voici les principales propriétés de ces champs modulaires :

- D'associer à tout chtouca son pôle et son zéro définit un morphisme

$$(\infty, 0) : \mathrm{Cht}^r \to X \times X \qquad [\text{resp. } \mathrm{Cht}^r_N \to (X - N) \times (X - N)]$$

qui est lisse de dimension relative $2r - 2$.
- Chaque Cht^r_N est représentable fini étale galoisien de groupe $\mathrm{GL}_r(\mathcal{O}_N)$ au-dessus de $\mathrm{Cht}^r \times_{X \times X} (X - N) \times (X - N)$.
- Chaque Cht^r_N est muni d'une action du groupe $F^\times \backslash \mathbb{A}^\times$ et d'une action par correspondances finies étales de la sous-algèbre de Hecke $\mathcal{H}^r_N$.

On s'intéresse aux espaces de cohomologie ℓ-adique à supports compacts $H^\nu_c(\mathrm{Cht}^r_N / a^{\mathbb{Z}})$, $0 \leq \nu \leq 2(2r-2)$, des $\mathrm{Cht}^r_N / a^{\mathbb{Z}}$ au-dessus du point générique $\mathrm{Spec}\, F^2$ de $X \times X$. Ils sont munis d'une double action du groupe de Galois G_{F^2} et des algèbres $\mathcal{H}^r_N$. Notre but va être de les calculer au moins partiellement et de montrer qu'ils réalisent l'application $\pi \mapsto \sigma_\pi$.

Chaque $\mathrm{Cht}^r_N / a^{\mathbb{Z}}$ n'a qu'un nombre fini de composantes connexes (ce qui correspond au fait que le volume de $\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A}) / a^{\mathbb{Z}}$ est fini) mais il n'est pas de type fini (de même que $\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A}) / a^{\mathbb{Z}}$ n'est pas compact), ses espaces de cohomologie ℓ-adique sont de dimension infinie (de même que la décomposition spectrale de $L^2(\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A}) / a^{\mathbb{Z}})$ fait apparaître des sommes continues) et les correspondances de Hecke composées avec les éléments de Frobenius ont une infinité de points fixes (de même que les intégrales "$\mathrm{Tr}(h)$" divergent).

Pour surmonter cette difficulté, on considère à nouveau un polygone de troncature $p : [0, r] \to \mathbb{R}_+$ et on demande aux chtoucas que leur polygone canonique de Harder-Narasimhan (qui se définit facilement car un chtouca est un fibré muni d'une structure supplémentaire) soit $\leq p$. Cela définit des ouverts $\mathrm{Cht}^{r, \overline{p} \leq p}_N$ dans les Cht^r_N dont les quotients par $a^{\mathbb{Z}}$ sont de type fini (de même que les intégrales $\mathrm{Tr}^{\leq p}(h)$ convergent).

Une fonction $h \in \mathcal{H}_N^r$ contient en facteur la fonction caractéristique $\mathbb{1}_y$ de $\mathrm{GL}_r(O_y)$ dans $\mathrm{GL}_r(F_y)$ en presque tout $y \in |X|$. On considère un point fermé x de $(X-N)\times(X-N)$ dont les deux projections $\infty, 0 \in |X|$ sont distinctes et vérifient cette propriété, et un multiple $s = \deg(\infty)\, s' = \deg(0)\, u'$ de $\deg(x)$. Le composé $h \times \mathrm{Frob}^s$ a dans la fibre de $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ au-dessus de n'importe quel point de $(X\times X)(\overline{\mathbb{F}}_q)$ supporté par x un nombre fini de points fixes noté

$$\mathrm{Lef}_x(h\times \mathrm{Frob}^s, \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}})\,.$$

En utilisant une description adélique des points fixes (rendue possible par la proximité des chtoucas avec le quotient $\mathrm{GL}_r(F)\backslash \mathrm{GL}_r(\mathbb{A})$) et le calcul des intégrales orbitales des fonctions sphériques $h_\infty^{-s'}$ et $h_0^{u'}$ sur $\mathrm{GL}_r(F_\infty)$ et $\mathrm{GL}_r(F_0)$ dont les transformés de Satake sont les polynômes symétriques $q^{s\frac{(r-1)}{2}}(Z_1^{-s'}+\cdots+Z_r^{-s'})$ et $q^{s\frac{(r-1)}{2}}(Z_1^{u'}+\cdots+Z_r^{u'})$, on relie ce nombre à la trace tronquée $\mathrm{Tr}^{\leq p}(h')$ de la fonction h' déduite de h en remplaçant les facteurs $\mathbb{1}_\infty$ et $\mathbb{1}_0$ par $h_\infty^{-s'}$ et $h_0^{u'}$. Par combinaison avec la formule des traces d'Arthur-Selberg, on obtient :

Théorème (comptage des points fixes). – *Dans la situation ci-dessus et si p est assez convexe et* $\deg(\infty)$, $\deg(0)$, *s sont assez grands en fonction de h, on a*

$$\begin{aligned}
&\frac{1}{r!}\sum_{k=1}^{r!}\mathrm{Lef}_{(\mathrm{Frob}_X^k\times \mathrm{Id}_X)(x)}(h\times\mathrm{Frob}^s, \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}})\\
=\;& q^{(r-1)s}\sum_{\substack{\pi\in\{\pi\}_r\\ \chi_\pi(a)=1}}\mathrm{Tr}_\pi(h)(z_1(\pi_\infty)^{-s'}+\cdots+z_r(\pi_\infty)^{-s'})\\
&\qquad\qquad (z_1(\pi_0)^{u'}+\cdots+z_r(\pi_0)^{u'})\\
+\;& \textit{autres termes où apparaissent les valeurs propres de Hecke en } \infty \textit{ et } 0 \textit{ des}\\
&\textit{représentations automorphes cuspidales des } \mathrm{GL}_{r_1}\times\cdots\times\mathrm{GL}_{r_k},\\
&r_1+\cdots+r_k=r.
\end{aligned}$$

Si dans cette formule il n'y avait que le terme principal et si on disposait sur $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ d'un théorème des points fixes de Grothendieck-Lefschetz interprétant ces nombres comme la trace sur la cohomologie d'une action des $h\times\mathrm{Frob}_x^{-s/\deg(x)}$ (en notant $\mathrm{Frob}_x \in G_{F^2}$ un élément de Frobenius en x), on aurait une représentation de $\mathcal{H}_N^r \times G_{F^2}$ qui se décomposerait nécessairement en

$$\bigoplus_{\substack{\pi\in\{\pi\}_r\\ \chi_\pi(a)=1}}(\pi\cdot\mathbb{1}_N)\otimes[\sigma_\pi\boxtimes\check{\sigma}_\pi]\,(1-r)$$

et on aurait construit l'application cherchée $\pi\mapsto\sigma_\pi$.

Mais le terme principal n'est pas seul, il y a aussi des termes complémentaires qui dépendent de p et font apparaître les représentations automorphes cuspidales en rangs $<r$, et surtout il n'y a même pas d'action de $\mathcal{H}_N^r$ car les correspondances de Hecke de $\mathrm{Cht}_N^r/a^{\mathbb{Z}}$ ne stabilisent pas les ouverts $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ (ce qui correspond au fait que les traces tronquées $\mathrm{Tr}^{\leq p}(h)$ ne sont pas invariantes par conjugaison).

3. Donner un sens cohomologique au comptage des points fixes

On vient de dire que les ouverts $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ ne sont pas stabilisés par les correspondances de Hecke si bien qu'en dehors du cas $h = \mathbb{1}_N$ (c'est-à-dire de la seule action de G_{F^2}) les nombres $\mathrm{Lef}_x(h \times \mathrm{Frob}^s, \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}})$ n'ont pas a priori de sens cohomologique. On va voir qu'en fait il est possible de leur en donner un plus élaboré.

3.1. Compactifications

On compactifie les ouverts tronqués $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ pour retrouver des correspondances de Hecke qui agissent sur la cohomologie. On commence par le cas sans niveau :

Pour tout polygone de troncature p assez convexe en fonction du genre de la courbe X, on construit un champ $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}}$ algébrique au sens d'Artin, dont les groupes d'automorphismes sont finis (mais ramifiés), qui est muni d'une action de $F^\times\backslash\mathbb{A}^\times$ et d'un morphisme lisse sur $X \times X$, qui contient $\mathrm{Cht}^{r,\overline{p}\leq p}$ comme ouvert, dont le bord est un diviseur à croisements normaux relatif et enfin dont le quotient par $a^{\mathbb{Z}}$ est propre (en particulier de type fini et séparé) sur $X \times X$. Les strates de bord $\mathrm{Cht}_{\underline{r}}^{r,\overline{p}\leq p}$ sont naturellement indexées par les partitions $\underline{r} = (r = r_1 + \cdots + r_k)$ de l'entier r. Si on distingue le degré d des chtoucas avec donc $\mathrm{Cht}^r = \coprod_{d\in\mathbb{Z}} \mathrm{Cht}^{r,d}$, $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}} = \coprod_{d\in\mathbb{Z}} \overline{\mathrm{Cht}^{r,d,\overline{p}\leq p}}$ et $\mathrm{Cht}_{\underline{r}}^{r,\overline{p}\leq p} = \coprod_{d\in\mathbb{Z}} \mathrm{Cht}_{\underline{r}}^{r,d,\overline{p}\leq p}$, chaque $\mathrm{Cht}_{\underline{r}}^{r,d,\overline{p}\leq p}$ est essentiellement de la forme

$$\mathrm{Cht}^{r_1,d_1,\overline{p}\leq p_1} \times_X \mathrm{Cht}^{r_2,d_2,\overline{p}\leq p_2} \times_{X,\mathrm{Frob}} \cdots \times_{X,\mathrm{Frob}} \mathrm{Cht}^{r_k,d_k,\overline{p}\leq p_k}$$

où $d_1, \ldots, d_k$ et $p_1, \ldots, p_k$ sont des degrés et polygones de troncatures qui se déduisent de d et p. La strate $\mathrm{Cht}_{\underline{r}}^{r,\overline{p}\leq p}$ est munie d'un morphisme vers $X \times X^{k-1} \times X$ où le premier facteur X est le pôle, le dernier facteur X est le zéro et les $k-1$ facteurs X supplémentaires sont appelés les dégénérateurs.

Toute fonction $h \in \mathcal{H}_{\emptyset}^r$ induit une correspondance sur $\mathrm{Cht}^r/a^{\mathbb{Z}}$. On peut considérer sa trace dans $(\mathrm{Cht}^{r,\overline{p}\leq p}/a^{\mathbb{Z}})^2$ puis la normalisation de celle-ci sur $(\overline{\mathrm{Cht}^{r,\overline{p}\leq p}}/a^{\mathbb{Z}})^2$. Comme $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}}/a^{\mathbb{Z}}$ est propre et lisse sur $X \times X$, elle agit sur sa cohomologie.

Considérons maintenant un niveau $N \hookrightarrow X$. La normalisation $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}/a^{\mathbb{Z}}$ de $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}}/a^{\mathbb{Z}} \times_{X\times X} (X-N)\times(X-N)$ dans $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ contient $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ comme ouvert et elle est propre sur $(X-N)\times(X-N)$ mais elle n'est pas lisse et l'auteur ignore comment résoudre ses singularités. Toutefois, en demandant que non seulement le pôle et le zéro mais aussi les dégénérateurs évitent N, on définit un ouvert $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$ de $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}/a^{\mathbb{Z}}$ qui contient strictement $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$, est lisse sur $X \times X$ et dont le bord est un diviseur à croisements normaux relatif. Les

correspondances de Hecke $h \in \mathcal{H}_N^r$ agissent sur sa cohomologie d'après le théorème suivant :

Théorème de stabilité globale. – *Les correspondances de Hecke étendues par normalisation sur* $(\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}})^2$ *stabilisent* $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$ *au sens que leurs deux projections sur celui-ci sont propres.*

Ce résultat correspond sans doute au phénomène suivant dans la formule des traces d'Arthur-Selberg : Le défaut d'invariance par conjugaison des traces tronquées $\mathrm{Tr}^{\leq p}(h)$ se mesure par les dérivées logarithmiques des opérateurs d'entrelacement de Langlands. Or ces opérateurs sont des produits eulériens et leurs dérivées logarithmiques sont des sommes sur tous les points $x \in |X|$ lesquels correspondent exactement aux valeurs possibles des dégénérateurs. Il n'y a pas d'entrelacement donc pas d'instabilité d'un point de $|X|$ à un autre et en demandant au pôle, au zéro et aux dégénérateurs d'éviter certains points de $|X|$, on garde la propriété de stabilité globale que vérifie automatiquement la compactification toute entière $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}/a^{\mathbb{Z}}$.

3.2. Cohomologie négligeable et cohomologie essentielle

En résumé, on a maintenant toutes les structures et informations qu'on pourrait souhaiter mais elles sont dispersées entre les différents objets $\mathrm{Cht}_N^r/a^{\mathbb{Z}}$, $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ et $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$:

Sur $\mathrm{Cht}_N^r/a^{\mathbb{Z}}$, on a une action de l'algèbre de Hecke $\mathcal{H}_N^r$ mais la cohomologie est de dimension infinie et les ensembles de points fixes sont infinis.

L'ouvert $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ est de type fini et on y a une formule de comptage des points fixes qui s'exprime en termes automorphes mais on a perdu l'action des correspondances de Hecke.

Enfin, sur $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$ et sa cohomologie il y a à nouveau une action de chaque $h \in \mathcal{H}_N^r$ mais on ne donne pas de comptage des points fixes et surtout il n'y a pas d'action de l'algèbre $\mathcal{H}_N^r$ car la normalisation des correspondances de Hecke ne commute pas avec la multiplication.

L'idée est de définir dans toute représentation ℓ-adique de G_{F^2} (après semi-simplification) une partie "négligeable" et une partie "essentielle", de façon que les $\mathrm{Cht}_N^r/a^{\mathbb{Z}}$, $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ et $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$ aient même cohomologie essentielle (ce qui permettra de rassembler sur celle-ci les informations dispersées dont on dispose) et que les traces des actions sur la cohomologie essentielle soient données par les termes principaux dans la formule de comptage des points fixes, ceux associés aux représentations automorphes cuspidales $\pi \in \{\pi\}_r$ de GL_r.

Or on s'attend à ce que les $\pi \in \{\pi\}_r$ correspondent à des représentations ℓ-adiques de G_F irréductibles de dimension r. Quant aux termes complémentaires dans la formule de comptage, ils sont associés aux représentations automorphes cuspidales en rangs $< r$ lesquelles correspondent, d'après l'hypothèse de récurrence, aux représentations ℓ-adiques de G_F irréductibles de dimension $< r$. Cela dicte la

définition suivante (où $q', q'' : G_{F^2} \rightrightarrows G_F$ sont les deux homomorphismes induits par $X \times X \rightrightarrows X$) :

Définition. – *Une représentation ℓ-adique irréductible de G_{F^2} est dite "r-négligeable" si elle est facteur direct d'une représentation de la forme*

$$q'^{*}\,\sigma' \otimes q''^{*}\,\sigma''$$

avec σ', σ'' deux représentations de G_F irréductibles de rangs $< r$.

Elle est dite "essentielle" sinon.

3.3. Séparation et identification de la cohomologie essentielle

Dans un premier temps, on oublie complètement les correspondances de Hecke et on ne considère que l'action de G_{F^2} sur les différents espaces de cohomologie ℓ-adique à supports compacts. On montre :

Proposition. – *Pour tout niveau $N \hookrightarrow X$, la cohomologie de $\mathrm{Cht}^{r}_{N}/a^{\mathbb{Z}}$ et des $\mathrm{Cht}^{r,\overline{p}\leq p}_{N}/a^{\mathbb{Z}}$ et $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}_{N}}'/a^{\mathbb{Z}}$ (et la cohomologie d'intersection des $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}_{N}}/a^{\mathbb{Z}}$ si $N \neq \emptyset$) ont la même partie essentielle H^{ess}_{N}. Elle est concentrée en degré médian $\nu = 2r-2$ et pure de poids $2r-2$. Si x est un point fermé de $(X-N)\times(X-N)$ dont les deux projections $\infty, 0 \in |X|$ sont distinctes et $s = \deg(\infty)\,s' = \deg(0)\,u'$ est un multiple de $\deg(x)$, on a*

$$\mathrm{Tr}_{H^{\mathrm{ess}}_{N}}(\mathrm{Frob}_{x}^{-s/\deg(x)}) \;=\; q^{(r-1)s} \sum_{\substack{\pi \in \{\pi\}_r \\ \chi_\pi(a)=1}} \dim(\pi \cdot \mathbb{1}_N)(z_1(\pi_\infty)^{-s'} + \cdots + z_r(\pi_\infty)^{-s'})$$
$$(z_1(\pi_0)^{u'} + \cdots + z_r(\pi_0)^{u'})\,.$$

Quand $N = \emptyset$, la cohomologie du bord $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}}/a^{\mathbb{Z}} - \mathrm{Cht}^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ est r-négligeable car il est réunion de strates $\mathrm{Cht}^{r,\overline{p}\leq p}_{\underline{r}}/a^{\mathbb{Z}}$ indexées par les partitions $\underline{r} = (r = r_1 + \cdots + r_k)$, $k \geq 2$, qui se dévissent en termes de $\mathrm{Cht}^{r_1}, \ldots, \mathrm{Cht}^{r_k}$. Pour $N \neq \emptyset$, un argument plus sophistiqué utilisant ce dévissage montre aussi que les différences entre $\mathrm{Cht}^{r,\overline{p}\leq p}_{N}/a^{\mathbb{Z}}$, $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}_{N}}'/a^{\mathbb{Z}}$ et $\overline{\mathrm{Cht}^{r,\overline{p}\leq p}_{N}}/a^{\mathbb{Z}}$ sont r-négligeables.

La formule de comptage des points fixes du paragraphe 2.4 (avec $h = \mathbb{1}_N$) et le théorème des points fixes de Grothendieck-Lefschetz donnent les traces des éléments $\mathrm{Frob}_{x}^{-s/\deg(x)}$ agissant sur la cohomologie des $\mathrm{Cht}^{r,\overline{p}\leq p}_{N}/a^{\mathbb{Z}}$. La séparation de la partie essentielle se fait en "testant" ces espaces de cohomologie contre des représentations irréductibles r-négligeables arbitraires et en regardant les pôles des fonctions L de paires obtenues. Cela utilise la correspondance de Langlands déjà connue en rangs $< r$ par hypothèse de récurrence et les propriétés classiques des fonctions L de paires tant du côté automorphe que galoisien (en particulier l'interprétation cohomologique de Grothendieck et le théorème de pureté de Deligne).

Enfin, $\mathrm{Cht}^{r}_{N}/a^{\mathbb{Z}}$ et les $\mathrm{Cht}^{r,\overline{p}\leq p}_{N}/a^{\mathbb{Z}}$ ont la même cohomologie essentielle car la formule obtenue pour les traces des $\mathrm{Frob}_{x}^{-s/\deg(x)}$ ne dépend pas de p.

3.4. Action et traces des correspondances de Hecke

Dans un second temps, on revient aux correspondances de Hecke $h \in \mathcal{H}_N^r$.

On met d'abord une action naturelle de l'algèbre $\mathcal{H}_N^r$ sur la cohomologie essentielle H_N^{ess} en prouvant que $H = H_c^{2r-2}(\mathrm{Cht}_N^r/a^{\mathbb{Z}}) = \varinjlim_p H_c^{2r-2}(\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}})$ admet une filtration finie $0 = H_0 \subsetneq \ldots \subsetneq H_i \subsetneq \ldots \subsetneq H_k = H$ respectée par la double action de G_{F^2} et de $\mathcal{H}_N^r$ dont les gradués impairs H_{2i+1}/H_{2i} sont entièrement r-négligeables et les gradués pairs H_{2i+2}/H_{2i+1} sont entièrement essentiels.

Pour terminer, il reste à montrer que pour $h \in \mathcal{H}_N^r$, x et $s = \deg(\infty)\, s' = \deg(0)\, u'$ comme dans l'énoncé de la formule de comptage, on a

$$\mathrm{Tr}_{H_N^{\mathrm{ess}}}(h \times \mathrm{Frob}_x^{-s/\deg(x)}) \;=\; q^{(r-1)s} \sum_{\substack{\pi \in \{\pi\}_r \\ \chi_\pi(a)=1}} \mathrm{Tr}_\pi(h)(z_1(\pi_\infty)^{-s'} + \cdots + z_r(\pi_\infty)^{-s'})\,(z_1(\pi_0)^{u'} + \cdots + z_r(\pi_0)^{u'})\,.$$

Pour cela, on a besoin d'une formule de Grothendieck-Lefschetz qui relie les nombres de points fixes $\mathrm{Lef}_x(h \times \mathrm{Frob}^s, \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}})$ à l'action de $h \times \mathrm{Frob}^s$ sur la cohomologie de $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$ (et ce sera suffisant car on n'a plus alors qu'à combiner une telle formule avec la formule de comptage et à identifier ce qui est "essentiel" des deux côtés par les arguments de fonctions L de paires).

On prouve en fait que pour toute $h \in \mathcal{H}_N^r$ fixée, il existe des correspondances cohomologiques $\mathrm{cl}(h)_{\underline{r}}$ agissant sur la cohomologie des strates $\overline{\mathrm{Cht}_{N,\underline{r}}^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$ du bord $\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}} - \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ telles que pour tout x et tout s comme plus haut on ait (en notant $H_c^*(\cdot) = \sum_\nu (-1)^\nu H_c^\nu(\cdot)^{\mathrm{ss}}$) :

$$\mathrm{Lef}_x(h \times \mathrm{Frob}_s, \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}) = \mathrm{Tr}_{H_c^*(\overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}})}(h \times \mathrm{Frob}_x^{-s/\deg(x)})$$
$$+ \sum_{\substack{\underline{r}=(r=r_1+\cdots+r_k) \\ k\geq 2}} (-1)^{k-1}\, \mathrm{Tr}_{H_c^*(\overline{\mathrm{Cht}_{N,\underline{r}}^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}})}(\mathrm{cl}(h)_{\underline{r}} \times \mathrm{Frob}_x^{-s/\deg(x)})\,.$$

La preuve de cette formule (qui justifie a posteriori le fait étrange qu'il y ait une formule de comptage des points fixes dans l'ouvert non stable $\mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$) repose sur la propriété géométrique suivante des correspondances de Hecke (qui paraît donc liée à l'existence même de la formule des traces d'Arthur-Selberg) :

Théorème de "stabilité locale". – *Dans $\mathfrak{X} = \overline{\mathrm{Cht}_N^{r,\overline{p}\leq p}}'/a^{\mathbb{Z}}$, les correspondances de Hecke h stabilisent l'ouvert $\mathfrak{X}_\emptyset = \mathrm{Cht}_N^{r,\overline{p}\leq p}/a^{\mathbb{Z}}$ "localement au voisinage de leurs points fixes".*

Cela signifie que si $\Gamma \xhookrightarrow{(p'_\Gamma, p''_\Gamma)} \mathfrak{X} \times \mathfrak{X}$ est le cycle qui supporte h, il existe un ouvert $U \subset \mathfrak{X} \times \mathfrak{X}$ contenant les points fixes de toutes les correspondances $\Gamma \times \mathrm{Frob}^n$, $n \in \mathbb{N}$, tel que

$$p'^{-1}_\Gamma(\mathfrak{X}_\emptyset) \cap U \subseteq p''^{-1}_\Gamma(\mathfrak{X}_\emptyset) \cap U\,.$$

4. Applications du théorème

La correspondance de Langlands entre représentations ℓ-adiques irréductibles du groupe de Galois G_F de F et représentations automorphes cuspidales des groupes GL_r sur F a des conséquences immédiates et importantes dans les deux sens. Voici les principales :

4.1. Conséquences sur les faisceaux ℓ-adiques

On a d'abord des conséquences très fortes dans le cas des courbes :

Théorème. – *Soient X' un ouvert de la courbe X et σ un faisceau ℓ-adique lisse sur X', qui est irréductible de rang r et dont le déterminant est un caractère d'ordre fini. Alors :*

(i) *Il existe un corps de nombres $E \subset \overline{\mathbb{Q}}$ tel qu'en tout point fermé $x \in |X'|$, le polynôme $\mathrm{L}_x(\sigma, T)^{-1} = \det_\sigma(1 - T \cdot \mathrm{Frob}_x^{-1})$ soit à coefficients dans E.*

(ii) *En tout $x \in |X'|$, les racines du polynôme $\det_\sigma(1 - T \cdot \mathrm{Frob}_x^{-1})$ sont des nombres algébriques dont toutes les images complexes sont de module 1. Ce sont des unités λ-adiques en toutes les places λ non archimédiennes et premières à q de E.*

(iii) *Pour toute place λ de E au-dessus d'un nombre premier ℓ' ne divisant pas q, il existe sur X' un faisceau ℓ'-adique σ_λ lisse et irréductible de rang r tel que*

$$\det_\sigma(1 - T \cdot \mathrm{Frob}_x^{-1}) = \det_{\sigma_\lambda}(1 - T \cdot \mathrm{Frob}_x^{-1}), \qquad \forall\, x \in |X'|\,.$$

On voit qu'en dimension 1 sur $\mathbb{F}_q$, un faisceau ℓ-adique irréductible dont le déterminant est d'ordre fini est "pur de poids 0" et "il ne dépend pas du choix de ℓ". La première de ces deux propriétés s'étend automatiquement en dimension arbitraire :

Corollaire. – *Soient $\mathfrak{X}$ une variété normale de type fini sur $\mathbb{F}_q$ et σ un faisceau ℓ-adique lisse sur $\mathfrak{X}$ qui est irréductible et dont le déterminant est d'ordre fini.*

Alors en tout point fermé x de $\mathfrak{X}$, les valeurs propres de Frobenius de σ sont des nombres algébriques dont toutes les images complexes sont de modules 1.

Et ce sont des unités λ-adiques pour toute place λ première à q.

4.2. Conséquences sur les représentations automorphes

Pour les représentations automorphes cuspidales des groupes linéaires sur $F = F(X)$, on a d'abord des conséquences sur les valeurs propres de Hecke et les fonctions L de paires :

Théorème. – (i) (Conjecture de Ramanujan-Petersson) *Pour toute $\pi \in \{\pi\}_r$, les facteurs locaux π_x de π en les $x \in |X|$ sont tempérés.*

En particulier, en les x où π_x est non ramifié, ses valeurs propres de Hecke vérifient

$$|z_i(\pi_x)| = 1\,, \qquad 1 \leq i \leq r\,.$$

(ii) (Hypothèse de Riemann généralisée) *Pour toute paire* $\pi \in \{\pi\}_r$, $\pi' \in \{\pi\}_{r'}$, *tous les zéros de la fonction* L *globale*

$$\mathrm{L}(\pi \times \pi', T)$$

sont sur le cercle

$$|T| = q^{-1/2}\,.$$

La partie (ii) du théorème est la traduction en termes automorphes du théorème de pureté de Deligne.

D'autre part, on a les cas particuliers suivants de la fonctorialité de Langlands (où, pour toute extension finie F' de F, $\mathbb{A}_{F'}$ désigne son anneau des adèles et $X_{F'}$ la courbe projective lisse associée qui est un revêtement fini de $X = X_F$) :

Théorème. – (i) (Existence du produit tensoriel automorphe) *Soient* π *et* π' *deux représentations automorphes cuspidales de* $\mathrm{GL}_r(\mathbb{A}_F)$ *et* $\mathrm{GL}_{r'}(\mathbb{A}_F)$. *Alors il existe une partition* $rr' = r_1 + \cdots + r_k$ *et des représentations automorphes cuspidales* $\pi^1, \ldots, \pi^k$ *de* $\mathrm{GL}_{r_1}(\mathbb{A}_F), \ldots, \mathrm{GL}_{r_k}(\mathbb{A}_F)$ *qui, en tout* $x \in |X_F|$ *où* π *et* π' *sont non ramifiées, sont elles-mêmes non ramifiées et vérifient*

$$\{z_j(\pi_x)\, z_{j'}(\pi'_x) \mid 1 \leq j \leq r\,,\ 1 \leq j' \leq r'\} = \coprod_{1 \leq i \leq k} \{z_1(\pi^i_x), \ldots, z_{r_i}(\pi^i_x)\}$$

(ii) (Changement de base) *Soient* F' *une extension finie de* F *et* π *une représentation automorphe cuspidale de* $\mathrm{GL}_r(\mathbb{A}_F)$. *Alors il existe une partition* $r = r_1 + \cdots + r_k$ *et des représentations automorphes cuspidales* ${\pi'}^1, \ldots, {\pi'}^k$ *de* $\mathrm{GL}_{r_1}(\mathbb{A}_{F'}), \ldots, \mathrm{GL}_{r_k}(\mathbb{A}_{F'})$, *non ramifiées en tout point* $x' \in |X_{F'}|$ *de degré* $\frac{\deg(x')}{\deg(x)}$ *au-dessus d'un point* $x \in |X_F|$ *où* π *est non ramifiée et qui vérifient*

$$\left\{z_1(\pi_x)^{\frac{\deg(x')}{\deg(x)}}, \ldots, z_r(\pi_x)^{\frac{\deg(x')}{\deg(x)}}\right\} = \coprod_{1 \leq i \leq k} \{z_1({\pi'}^i_x), \ldots, z_{r_i}({\pi'}^i_x)\}\,.$$

(iii) (Induction automorphe) *Soient* F' *une extension de* F *de degré* d *et* π' *une représentation automorphe cuspidale de* $\mathrm{GL}_r(\mathbb{A}_{F'})$. *Alors il existe une partition* $rd = r_1 + \cdots + r_k$ *et des représentations automorphes cuspidales* $\pi^1, \ldots, \pi^k$ *de* $\mathrm{GL}_{r_1}(\mathbb{A}_F), \ldots, \mathrm{GL}_{r_k}(\mathbb{A}_F)$, *non ramifiées en tout point* $x \in |X_F|$ *au-dessus duquel les points* $x' \in |X_{F'}|$ *sont non ramifiées sur* x *et pour* π' *et qui vérifient*

$$\prod_{1 \leq i \leq k} \mathrm{L}_x(\pi^i, T) = \prod_{x'|x} \mathrm{L}_{x'}(\pi', T)\,.$$

Bibliographie

[1] J.G. Arthur, A trace formula for reductive groups I: terms associated to classes in $G(\mathbb{Q})$, *Duke Math. J.* 45 (1978), 911–952.

[2] J.G. Arthur, A trace formula for reductive groups II: applications of a truncation operator, *Comp. Math.* 40 (1) (1980), 87–121.

[3] J.G. Arthur, The trace formula for noncompact quotient, *Proceedings of ICM*, Varsovie (1983), Vol. 2, 849–859.
[4] J.W. Cogdell et I.I. Piatetski-Shapiro, Converse theorems for GL_n, *Pub. Math. IHES* 79 (1994), 157–214.
[5] P. Deligne, La conjecture de Weil II, *Publ. Math. IHES* 52 (1980), 137–252.
[6] V.G. Drinfeld, Langland's Conjecture for GL(2) over functional Fields, *Proceedings of ICM*, Helsinki (1978), 565–574.
[7] V.G. Drinfeld, Two-dimensional ℓ-adic representations of the fundamental group of a curve over a finite field and automorphic forms on GL(2), *Amer. J. Math.* 105 (1983), 85–114.
[8] V.G. Drinfeld, Varieties of modules of F-sheaves, *Funct. Anal. Appl.* 21 (1987), 107–122.
[9] V.G. Drinfeld, Cohomology of compactified manifolds of modules of F-sheaves of rank 2, *J. Sov. Math.* 46 (1989), 1789–1821.
[10] E. Frenkel, D. Gaitsgory et K. Vilonen, On the geometric Langlands conjecture, *J. AMS* 15 (2002), 367–417.
[11] D. Gaitsgory, On a vanishing conjecture appearing in the geometric Langlands correspondence, *prépublication*, http://arXiv.org/abs/math/0204081, (2002).
[12] L. Lafforgue, Chtoucas de Drinfeld et conjecture de Ramanujan-Petersson, *Astérisque* 243, SMF (1997).
[13] L. Lafforgue, Une compactification des champs classifiant les chtoucas de Drinfeld, *J. AMS* 11 (1998), 1001–1036.
[14] L. Lafforgue, Chtoucas de Drinfeld et correspondance de Langlands, *Inv. Math.* 147 (2002), 1–241.
[15] L. Lafforgue, Cours à l'Institut Tata sur les chtoucas de Drinfeld et la correspondance de Langlands, *prépublication*, http://www.ihes.fr/PREPRINTS/M02/Resu/resu-M02-45.html, à paraître dans le volume de l'année spéciale du TIFR de Bombay sur le programme de Langlands géométrique, (2002).
[16] R.P. Langlands, Problems in the theory of Automorphic Forms, dans *Lectures in Modern Analysis and Applications III, LNM* 170 (1970), 18–61.
[17] R.P. Langlands, Automorphic representations, Shimura varieties, and motives. Ein Märchen, *PSPM* XXXIII (2) (1979), 205–246.
[18] G. Laumon, Transformation de Fourier, constantes d'équations fonctionnelles et conjecture de Weil, *Pub. Math. IHES* 65 (1987), 131–210.
[19] G. Laumon, Cohomology of Drinfeld Modular Varieties, volumes I et II, Cambridge U.P. (1996).
[20] G. Laumon, La correspondance de Langlands sur les corps de fonctions, *Séminaire Bourbaki*, exposé 873 (mars 2000).
[21] G. Laumon, Travaux de Frenkel, Gaitsgory et Vilonen sur la correspondance de Drinfeld-Langlands, *Séminaire Bourbaki*, exposé 906 (juin 2002).
[22] G. Laumon, M. Rapoport et U. Stuhler, $\mathcal{D}$-elliptic sheaves and the Langlands correspondence, *Inv. Math.* 113 (1993), 217–338.
[23] C. Mœglin et J.-L. Waldspurger, Décomposition spectrale et séries d'Eisenstein, *Prog. in Math.* 113 (1989), Birkhaüser.

ICM 2002 · Vol. I · 401–422

Pattern Theory: The Mathematics of Perception

David Mumford*

Abstract

Is there a mathematical theory underlying intelligence? Control theory addresses the output side, motor control, but the work of the last 30 years has made clear that perception is a matter of Bayesian statistical inference, based on stochastic models of the signals delivered by our senses and the structures in the world producing them. We will start by sketching the simplest such model, the hidden Markov model for speech, and then go on illustrate the complications, mathematical issues and challenges that this has led to.

Keywords and Phrases: Perception, Speech, Vision, Bayesian, Statistics, Inference, Markov.

1. Introduction

How can we understand intelligent behavior? How can we design intelligent computers? These are questions that have been discussed by scientists and the public at large for over 50 years. As mathematicians, however, the question we want to ask is "is there a *mathematical* theory underlying intelligence?" I believe the first mathematical attack on these issues was Control Theory, led by Wiener and Pontryagin. They were studying how to design a controller which drives a motor affecting the world and also sits in a feedback loop receiving measurements from the world about the effect of the motor action. The goal was to control the motor so that the world, as measured, did something specific, i.e. move the tiller so that the boat stays on course. The main complication is that nothing is precisely predictable: the motor control is not exact, the world does unexpected things because of its complexities and the measurements you take of it are imprecise. All this led, in the simplest case, to a beautiful analysis known as the Wiener-Kalman-Bucy filter (to be described below).

But Control Theory is basically a theory of the output side of intelligence with the measurements modeled in the simplest possible way: e.g. linear functions of the

*Division of Applied Mathematics, Brown University, Providence RI 02912, USA. E-mail: David_Mumford@brown.edu

state of the world system being controlled plus additive noise. The real input side of intelligence is perception in a much broader sense, the analysis of all the noisy incomplete signals which you can pick up from the world through natural or artificial senses. Such signals typically display a mix of distinctive patterns which tend to repeat with many kinds of variations and which are confused by noisy distortions and extraneous clutter. The interesting and important structure of the world is thus coded in these signals, using a code which is complex but not perversely so.

1.1. Logic vs. Statistics

The first serious attack on problems of perception was the attempt to recognize speech which was launched by the US defense agency ARPA in 1970. At this point, there were two competing ideas of what was the right formalism for combining the various clues and features which the raw speech yielded. The first was to use logic or, more precisely, a set of 'production rules' to augment a growing database of true propositions about the situation at hand. This was often organized in a 'blackboard', a two-dimensional buffer with the time of the asserted proposition plotted along the x-axis and the level of abstraction (i.e. signal — phone — phoneme — syllable — word — sentence) along the y-axis. The second was to use statistics, that is, to compute probabilities and conditional probabilities of various possible events (like the identity of the phoneme being pronounced at some instant). These statistics were computed by what was called the 'forward-backward' algorithm, making 2 passes in time, before the final verdict about the most probable translation of the speech into words was found. This issue of logic vs. statistics in the modeling of thought has a long history going back to Aristotle about which I have written in [M].

I think it is fair to say that statistics won. People in speech were convinced in the 1970's, artificial intelligence researchers converted during the 1980's as expert systems needed statistics so clearly (see Pearl's influential book [P]), but vision researchers were not converted until the 1990's when computers became powerful enough to handle the much larger datasets and algorithms needed for dealing with 2D images.

The biggest reason why it is hard to accept that statistics underlies all our mental processes — perception, thinking and acting — is that we are not consciously aware of 99% of the ambiguities with which we deal every second. What philosophers call the 'raw qualia', the actual sensations received, do not make it to consciousness; what we are conscious of is a precise unambiguous enhancement of the sensory signal in which our expectations and our memories have been drawn upon to label and complete each element of the percept. A very good example of this comes from the psychophysical experiments of Warren & Warren [W] in 1970: they modified recorded speech by replacing a single phoneme in a sentence by a noise and played this to subjects. Remarkably, the subjects did *not* perceive that a phoneme was missing but believed they had heard the one phoneme which made the sentence semantically consistent:

Actual sound	Perceived words
the ¿eel is on the shoe	the *h*eel is on the shoe
the ¿eel is on the car	the *wh*eel is on the car
the ¿eel is on the table	the *m*eal is on the table
the ¿eel is on the orange	the *p*eel is on the orange

Two things should be noted. Firstly, this showed clearly that the actual auditory signal did not reach consciousness. Secondly, the choice of percept was a matter of probability, not certainty. That is, one might find some odd shoe with a wheel on it, a car with a meal on it, a table with a peel on it, etc. but the words which popped into consciousness were the most likely. An example from vision of a simple image, whose contents require major statistical reasoning to reconstruct, is shown in figure 1.

Figure 1: Why is this old man recognizable from a cursory glance? His outline threads a complex path amongst the cluttered background and is broken up by alternating highlights and shadows and by the wrinkles on his coat. There is no single part of this image which suggests a person unambiguously (the ear comes closest but the rest of his face can only be guessed at). No other object in the image stands out — the man's cap, for instance, could be virtually anything. Statistical methods, first grouping contours, secondly guessing at likely illumination effects and finally using probable models of clothes may draw him out. No known computer algorithm comes close to finding a man in this image.

It is important to clarify the role of probability in this approach. The uncertainty in a given situation need not be caused by observations of the world being truly unpredictable as in quantum mechanics or even effectively so as in chaotic phenomena. It is rather a matter of efficiency: in order to understand a sentence being spoken, we do not need to know all the things which affect the sound such as the exact acoustics of the room in which we are listening, nor are we even able to know other factors like the state of mind of the person we are listening to. In other words, we always have incomplete data about a situation. A vast number of physical

and mental processes are going on around us, some germane to the meaning of the signal, some extraneous and just cluttering up the environment. In this 'blooming, buzzing' world, as William James called it, we need to extract information and the best way to do it, apparently, is to make a stochastic model in which all the irrelevent events are given a simplified probability distribution. This is not unlike the stochastic approach to Navier-Stokes, where one seeks to replace turbulence or random molecular effects on small scales by stochastic perturbations.

1.2. The Bayesian setup

Having accepted that we need to use probabilities to combine bits and pieces of evidence, what is the mathematical set up for this? We need the following ingredients: a) a set of random variables, some of which describe the observed signal and some the 'hidden' variables describing the events and objects in the world which are causing this signal, b) a class of stochastic models which allow one to express the variability of the world and the noise present in the signals and c) specific parameters for the one stochastic model in this class which best describes the class of signals we are trying to decode now. More formally, we shall assume we have a set $\mathbf{x} = (\mathbf{x}_\text{o}, \mathbf{x}_\text{h})$ of observed and hidden random variables, which may have real values or discrete values in some finite or countable sets, we have a set $\boldsymbol{\theta}$ of parameters and we have a class of probability models $\Pr(\mathbf{x} \mid \boldsymbol{\theta})$ on the x's for each set of values of the $\boldsymbol{\theta}$'s. The crafting or learning of this model may be called the first problem in the mathematical theory of perception. It is usual to factor these probability distributions:

$$\Pr(\mathbf{x} \mid \boldsymbol{\theta}) = \Pr(\mathbf{x}_\text{o} \mid \mathbf{x}_\text{h}, \boldsymbol{\theta}) \cdot \Pr(\mathbf{x}_\text{h} \mid \boldsymbol{\theta}),$$

where the first factor, describing the likelihood of the observations from the hidden variables, is called the *imaging model* and the second, giving probabilities on the hidden variables, is called the *prior.* In the full Bayesian setting, one has an even stronger prior, a full probability model $\Pr(\mathbf{x}_h, \boldsymbol{\theta})$, including the parameters.

The second problem of perception is that we need to estimate the values of the parameters $\boldsymbol{\theta}$ which give the best stochastic model of this aspect of the world. This often means that you have some set of measurements $\{\mathbf{x}^{(\alpha)}\}$ and seek the value of $\boldsymbol{\theta}$ which maximizes their likelihood $\prod_\alpha \Pr(\mathbf{x}^{(\alpha)} \mid \boldsymbol{\theta})$. If the hidden variables as well as the observations are known, this is called supervised learning; if the hidden variables are not known, then it is unsupervised and one may maximize, for instance, $\prod_\alpha \sum_{\mathbf{x}_\text{h}} \Pr(\mathbf{x}_\text{o}^{(\alpha)}, \mathbf{x}_\text{h} \mid \boldsymbol{\theta})$. If one has a prior on the $\boldsymbol{\theta}$'s too, one can also estimate them from the mean or mode of the full posterior $\Pr(\boldsymbol{\theta} \mid \{\mathbf{x}^{(\alpha)}\})$.

Usually a more challenging problem is how many parameters $\boldsymbol{\theta}$ to include. At one extreme, there are simple 'off-the-shelf' models with very few parameters and, at the other extreme, there are fully non-parametric models with infinitely many parameters. Here the central issue is how much data one has: for any set of data, models with too few parameters distort the information the data contains and models with too many overfit the accidents of this data set. This is called the *bias-variance dilemma.* There are two main approaches to this issue. One is cross-validation: hold back parts of the data, train the model to have maximal likelihood

on the training set and test it by checking the likelihood of the held out data. There is also a beautiful theoretical analysis of the problem due principally to Vapnik [V] and involving the *VC dimension* of the models — the size of the largest set of data which can be split in all possible ways into more and less likely parts by different choices of $\boldsymbol{\theta}$.

As Grenander has emphasized, a very useful test for a class of models is to synthesize from it, i.e. choose random samples according to this probability measure and to see how well they resemble the signals we are accustomed to observing in the world. This is a stringent test as signals in the world usually express layers and layers of structure and the model tries to describe only a few of these.

The third problem of perception is using this machinary to actually perceive: we assume we have measured specific values $\mathbf{x}_o = \hat{\mathbf{x}}_o$ and want to infer the values of the hidden variables $\mathbf{x}_h$ in this situation. Given these observations, by Bayes' rule, the hidden variables are distributed by the so-called *posterior* distribution:

$$\Pr(\mathbf{x}_h \mid \hat{\mathbf{x}}_o, \boldsymbol{\theta}) = \frac{\Pr(\hat{\mathbf{x}}_o \mid \mathbf{x}_h, \boldsymbol{\theta}) \cdot \Pr(\mathbf{x}_h \mid \boldsymbol{\theta})}{\Pr(\hat{\mathbf{x}}_o \mid \boldsymbol{\theta})} \propto \Pr(\hat{\mathbf{x}}_o \mid \mathbf{x}_h, \boldsymbol{\theta}) \cdot \Pr(\mathbf{x}_h \mid \boldsymbol{\theta})$$

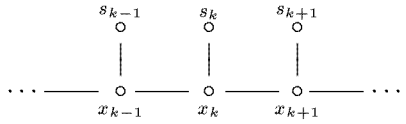
One may then want to estimate the mode of the posterior, the most likely value of $\mathbf{x}_h$. Or one may want to estimate the mean of some functions $f(\mathbf{x}_h)$ of the hidden variables. Or, if the posterior is often multi-modal and some evidence is expected to available later, one usually wants a more complete description or approximation to the full posterior distribution.

2. A basic example: HMM's and speech recognition

A convenient way to introduce the ideas of Pattern Theory is to outline the simple Hidden Markov Model method in speech recognition to illustrate many of the ideas and problems which occur almost everywhere. Here the observed random variables are the values of the sound signal $s(t)$, a pressure wave in air. The hidden random variables are the states of the speaker's mouth and throat and the identity of the phonemes being spoken at each instant. Usually this is simplified, replacing the signal by samples $s_k = s(k\Delta t)$ and taking for hidden variables a sequence x_k whose values indicate which phone in which phoneme is being pronounced at time $k\Delta t$. The stochastic model used is:

$$\Pr(x_\cdot, s_\cdot) = \prod_k p_1(x_k \mid x_{k-1}) p_2(s_k \mid x_k)$$

i.e. the $\{x_k\}$ form a Markov chain and each s_k depends only on x_k. This is expressed by the graph:



in which each variable corresponds to a vertex and the graphical Markov property holds: if 2 vertices a, b in the graph are separated by a subset S of vertices, then the variables associated to a and b are conditionally independent if we fix the variables associated to S.

This simple model works moderately well to decode speech because of the linear nature of the graph, which allows the ideas of dynamic programming to be used to solve for the marginal distributions and the modes of the hidden variables, given any observations $\hat{s}_\bullet$. This is expressed simply in the recursive formulas:

$$\begin{aligned}
\Pr(x_k \mid \hat{s}_{\leq k}) &= \frac{\sum_{x_{k-1}} p_1(x_k \mid x_{k-1}) p_2(\hat{s}_k \mid x_k) \Pr(x_{k-1} \mid \hat{s}_{\leq (k-1)})}{\sum_{x_k} \text{numerator}} \\
\operatorname{maxp}(x_k, \hat{s}_{\leq k}) &\underset{\text{def}}{=} \max_{x_{\leq (k-1)}} \Pr(x_k, x_{\leq k-1}, \hat{s}_{\leq k}) \\
&= \max_{x_{k-1}} \left(p_1(x_k \mid x_{k-1}) p_2(\hat{s}_x \mid x_k) \operatorname{maxp}(x_{k-1}, \hat{s}_{\leq (k-1)}) \right).
\end{aligned}$$

Note that if each x_k can take N values, the complexity of each time step is $O(N^2)$.

In any model, if you can calculate the conditional probabilities of the hidden variables and if the model is of exponential type, i.e.

$$\Pr(x_\bullet \mid \boldsymbol{\theta}_\bullet) = \frac{1}{Z(\boldsymbol{\theta})} e^{\sum_k \theta_k \cdot E_k(x_\bullet)},$$

then there is also an efficient method of optimizing the parameters $\boldsymbol{\theta}$. This is called the *EM algorithm* and, because it holds for HMM's, it is one of the key reasons for the early successes of the stochastic approach to speech recognition. For instance, a Markov chain $\{x_k\}$ is an exponential model if we let the $\boldsymbol{\theta}$'s be $\log(p(a \mid b))$ and write the chain probabilities as:

$$\Pr(x_\bullet) = e^{\sum_{a,b} \log(p(a|b)) |\{k | x_k = a, x_{k-1} = b\}|}.$$

The fundamental result on exponential models is that the $\boldsymbol{\theta}$'s are determined by the expectations $\hat{E}_k = \operatorname{Exp}(E_k)$ and that any set of expectations $\hat{E}_k$ that can be achieved in some probability model (with all probabilities non-zero), is also achieved in an exponential model.

2.1. Continuous and discrete variables

In this model, the observations s_k are naturally continuous random variables, like all primary measurements of the physical world. But the hidden variables are discrete: the set of phonemes, although somewhat variable from language to language, is always a small discrete set. This combination of discrete and continuous is characteristic of perception. It is certainly a psychophysical reality: for example experiments show that our perceptions lock onto one or another phoneme, resisting ambiguity (see [L], Ch.8, esp. p.176). But it shows itself more objectively in the low-level statistics of natural signals. Take almost any class of continuous real-valued signals $s(t)$ generated by the world and compile a histogram of their changes $x = s(t+\Delta t) - s(t)$ over some fixed time interval Δt. This empirical distribution will very

likely have kurtosis $(= \text{Exp}\left((x-\bar{x})^4\right)/\sigma(x)^4)$ greater than 3, the kurtosis of any Gaussian distribution! This means that, compared to a Gaussian distribution with the same mean and standard deviation, x has higher probability of being quite small or quite large but a lower probability of being average. Thus, compared to Brownian motion, $s(t)$ tends to move relatively little most of the time but to make quite large moves sometimes. This can be made precise by the theory of stochastic processes with iid increments, a natural first approximation to any stationary Markov process. The theory of such processes says that (a) their increments always have kurtosis at least 3, (b) if it equals 3 the process is Brownian and (c) if it is greater, samples from the process almost surely have discontinuities. At the risk of over-simplfying, we can say *kurtosis > 3 is nature's universal signal of the presence of discrete events/objects in continuous space-time.*

A classic example of this are stock market prices. Their changes (or better, changes in log(price)) have a highly non-Gaussian distribution with polynomial tails. In speech, the changes in the log(power) of the windowed Fourier transform show the same phenomenon, confirming that $s(t)$ cannot be decently modeled by colored Gaussian noise.

2.2. When compiling full probability tables is impractical

Applying HMM's in realistic settings, it usually happens that N is too large for an exhaustive search of complexity $O(N^2)$ or that the x_k are real valued and, when adequately sampled, again N is too large. There is one other situation in which the HMM-style approach works easily — the *Kalman filter.* In Kalman's setting, each variable x_k and s_k is real vector-valued instead of being discrete and p_1 and p_2 are *Gaussian distributions with fixed covariances and means depending linearly on the conditioning variable.* It is then easy to derive recursive update formulas, similar to those above, for the conditional distributions on each x_k, given the past data $\widehat{s}_{\leq k}$.

But usually, in the real-valued variable setting, the p's are more complex than Gaussian distributions. An example is the tracking problem in vision: the position and velocity x_k of some specific moving object at time $k\Delta t$ is to be inferred from a movie $\widehat{s}_k$, in which the object's location is confused by clutter and noise. It is clear that the search for the optimal reconstruction x_k must be pruned or approximated. A dramatic breakthrough in this and other complex situations has been to adapt the HMM/Kalman ideas by using weak approximations to the marginals $\Pr(x_k \mid \widehat{s}_{\leq k})$ by a finite set of samples, an idea called *particle filtering*:

$$\Pr(x_k \mid \widehat{s}_{\leq k}) \underset{\text{weak}}{\sim} \sum_{i=1}^{N} w_{i,k}\delta_{x_{i,k}}(x_k), \quad \text{that is,}$$

$$\text{Exp}(f(x_k) \mid \widehat{s}_{\leq k}) \approx \sum_{i=1}^{N} w_{i,k} f(x_{i,k}), \text{ for suitable } f.$$

This idea was proposed originally by Gordon, Salmond and Smith [G-S-S] and is developed at length in the recent survey [D-F-G]. An example with explicit estimates of the posterior from the work of Isard and Blake [I-B] is shown in figure 2. They

follow the version known as bootstrap particle filtering in which, for each k, N samples x_l' are drawn with replacement from the weak approximation above, each sample is propagated randomly to a new sample x_l'' at time $(k+1)$ using the prior $p(x_{k+1} \mid x_l')$ and these are reweighted proportional to $p(\widehat{s}_{k+1} \mid x_l'')$.

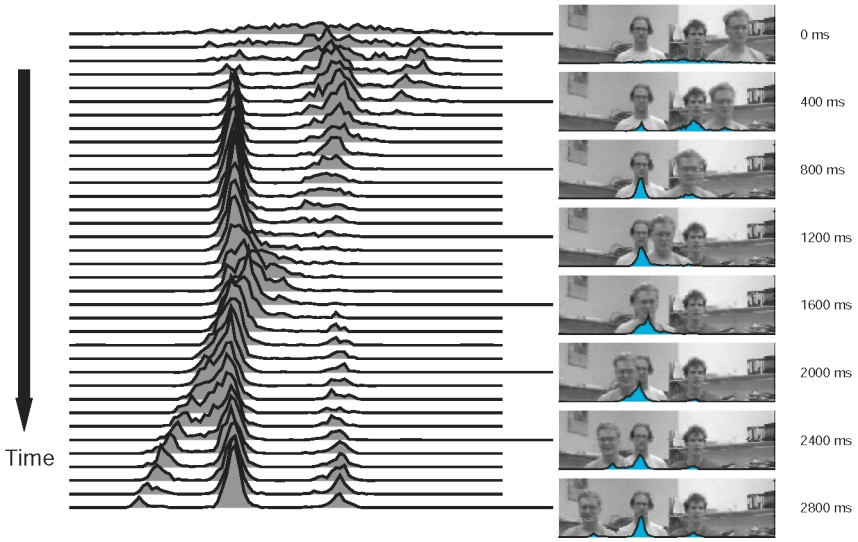


Figure 2: Work of Blake and Isard tracking three faces in a moving image sequence. The curves represent estimates of the posterior probability distributions for faces at each location obtained by smoothing the weighted sum of delta functions at the 'particles'. Note how multi-modal these are and how the tracker recovers from the temporary occlusion of one face by another.

2.3. No process in nature is truly Markov

A more serious problem with the HMM approach is that the Markov assumption is never really valid and it may be much too crude an approximation. Consider speech recognition. The finite lexicon of words clearly constrains the expected phoneme sequences, i.e. if x_k are the phonemes, then $p_1(x_k \mid x_{k-1})$ depends on the current word(s) containing these phonemes, i.e. on a short but variable part of the preceding string $\{x_{k-1}, x_{k-2}, \cdots\}$ of phonemes. To fix this, we could let x_k be a pair consisting of a word and a specific phoneme in this word; then $p_1(x_k \mid x_{k-1})$ would have two quite different values depending on whether x_{k-1} was the last phoneme in the word or not. Within a word, the chain needs only to take into account the variability with which the word can be pronounced. At word boundaries, it should use the conditional probabilities of word pairs. This builds much more of the patterns of the language into the model.

Why stop here? State-of-the-art speech recognizers go further and let x_k be

a pair of consecutive words plus a triphone[1] in the second word (or bridging the first and second word) whose middle phoneme is being pronounced at time $k\Delta t$. Then the transition probabilities in the HMM involve the statistics of 'trigrams', consecutive word triples in the language. But grammar tells us that words sequences are also structured into phrases and clauses of variable length forming a parse tree. These clearly affect the statistics. Semantics tells us that words sequences are further constrained by semantic plausibility ('sky' is more probable as the word following 'blue' than 'cry') and pragmatics tells us that sentences are part of human communications which further constrain probable word sequences.

All these effects make it clear that certain parts of the signal should be grouped together into units on a higher level and given labels which determine how likely they are to follow each other or combine in any way. This is the essence of grammar: higher order random variables are needed whose values are subsets of the low order random variables. The simplest class of stochastic models which incorporate variable length random substrings of the phoneme sequence are *probabilistic context free grammars* or PCFG's. Mathematically, they are a particular type of random branching tree.

Definition *A PCFG is a stochastic model in which the random variables are (a) a sequence of rooted trees* $\{\mathcal{T}_n\}$*, (b) a linearly ordered sequence of observations* s_k *and a 1:1 correspondence between the observations* s_k *and the leaves of the whole forest of trees such that the children of any vertex of any tree form an interval* $\{s_k, s_{k+1}, \cdots, s_{k'}\}$ *in time and (c) a set of labels* x_v *for each vertex. The probability model is given by conditional probabilities* $p_1(x_{v_k} \mid x_v)$ *for the labels of each child of each vertex*[2] *and* $p_2(s_k \mid x_{v_k})$ *for the observations, conditional on the label of the corresponding leaf.*

See figure 3 for an example. This has a Markov property if we define the 'extended' state x_k^* at leaf k to be not only the label x_k at this leaf but the whole sequence of labels on the path from this leaf to the root of the tree in which this leaf lies. Conditional on this state, the past and the future are independent.

This is a mathematically elegant and satisfying theory: unfortunately, it also fails, or rather explodes because, in carrying it out, the set of labels gets bigger and bigger. For instance, it is not enough to have a label for noun phrase which expands into an adjective plus a noun. The adjective and noun must agree in number and (in many languages) gender, a constraint that must be carried from the adjective to the noun (which need not be adjacent) via the label of the parent. So we need 4 labels, all combinations of singular/plural masculine/feminine noun phrases. And semantic constraints, such as Pr('blue sky') > Pr('blue cry'), would seem to require even more labels like 'colorable noun phrases'. Rather than letting the label set explode, it is better to consider a bigger class of grammars, which express these relations more succinctly but which are not so easily converted into HMM's: *unification grammars* [Sh] or *compositional grammars* [B-G-P]. The need for grammars of this type is

[1] So-called co-articulation effects mean that the pronunciation of a phoneme is affected by the preceding and suceeding phonemes.

[2] Caution to specialists: our label x_v is the name of the 'production rule' with this vertex as its head, esp. it fixes the arity of the vertex. We are doing it this way to simplify the Markov property.

especially clear when we look at formalisms for expressing the grouping laws in vision: see figure 3. The further development of stochastic compositional grammars, both in language and vision, is one of the main challenges today.

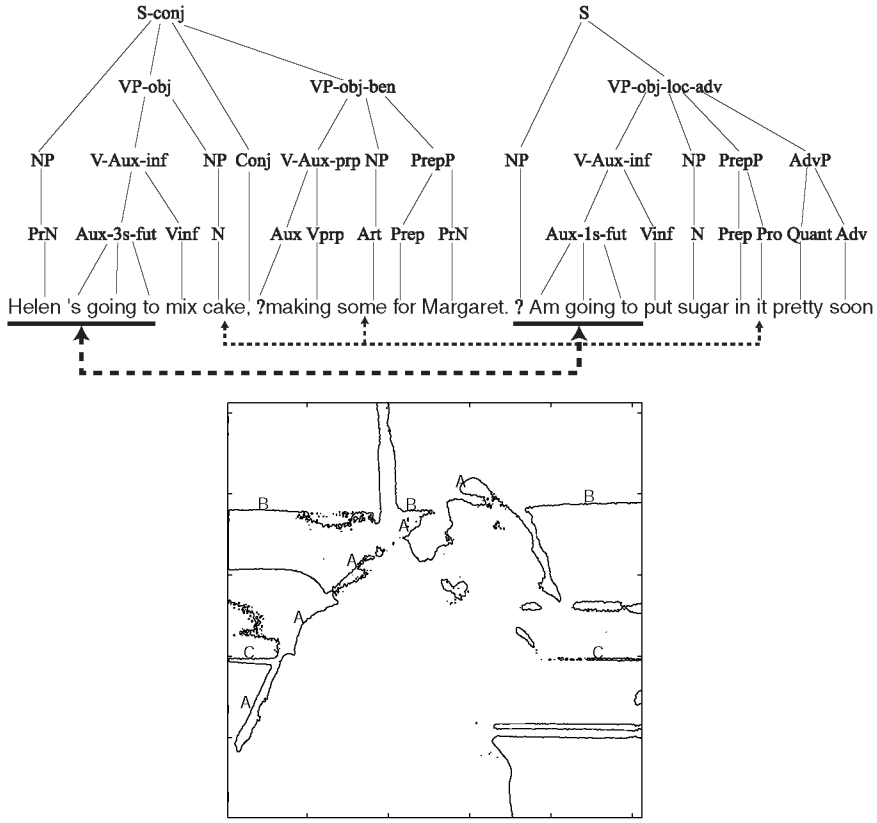


Figure 3: Grouping in language and vision: On top, parsing the not quite grammatical speech of a 2 1/2 year old Helen describing her own intentions ([H]): above the sentence, a context-free parse tree; below it, longer range non-Markov links — the identity 'cake'='some'='it' and the unification of the two parts 'Helen's going to' = '(I) am going to'. On the bottom, 2 kinds of grouping with an iso-intensity contour of the image in Figure 1: note the broken but visible contour of the back marked by 'A' and the occluded contours marked by 'B' and 'C' behind the man.

3. The 'natural degree of generality': MRF's or Graphical Models

The theory of HMM's deals with one-dimensional signals. But images, the signals occurring in vision, are usually two-dimensional — or three-dimensional for

MR scans and movies (3 space dimensions and 2 space plus 1 time dimension), even four-dimensional for echo cardiograms. On the other hand, the parse tree is a more abstract graphical structure and other 'signals', like medical data gathered about a particular patient, are structured in complex ways (e.g. a set of blood tests, a medical history). This leads to the basic insight of Grenander's Pattern Theory [G]: that the variables describing the structures in the world are typically related in a graphical fashion, edges connecting variables which have direct bearing on each other. Finding the right graph or class of graphs is a crucial step in setting up a satisfactory model for any type of patterns. Thus the applications, as well as the mathematical desire to find the most general setting for this theory, lead to the idea of replacing a simple chain of variables by a set of variables with a more general graphical structure. The general concept we need is that of a Markov random field:

Definition *A Markov random field is a graph $G = (V, E)$, a set of random variables $\{x_v\}_{v \in V}$, one for each vertex, and a joint probability distribution on these variables of the form:*

$$\Pr(x_\bullet) = \frac{1}{Z} e^{-\sum_C E_C(\{x_v\}_{v \in C})},$$

where C ranges over the cliques (fully connected subsets) of the graph, E_C are any functions and Z a constant. If the variables x_v are real-valued for $v \in V'$, we make this into a probability density, multiplying by $\prod_{n \in V'} dx_n$. Moreover, we can put each model in a family by introducing a temperature T and defining:

$$\Pr_T(x_\bullet) = \frac{1}{Z_T} e^{-\sum_C E_C(\{x_v\}_{v \in C})/T}.$$

These are also called *Gibbs models* in statistical mechanics (where the E_C are called *energies*) and *graphical models* in learning theory and, like Markov chains, are characterized by their conditional independence properties. This characterization, called the Hammersley-Clifford theorem, is that if two vertices $a, b \in V$ are separated by a subset $S \subset V$ (all paths in G from a to b must include some vertex in S), then x_a and x_b are conditionally independent given $\{x_v\}_{v \in S}$. The equivalence of these independence properties, plus the requirement that all probabilities be positive, with the simple explicit formula for the joint probabilities makes it very convincing that MRF's are a natural class of stochastic models.

3.1. The Ising model

This class of models is very expressive and many types of patterns which occur in the signals of nature can be captured by this sort of stochastic model. A basic example is the Ising model and its application to the image segmentation problem. In the simplest form, we take the graph G to be a square $N \times N$ grid with two layers, with observable random variables $p_{i,j} \in \mathbb{R}, 1 \leq i, j \leq N$ associated to the top layer and hidden random variables $x_{i,j} \in \{+1, -1\}$ associated to the bottom layer. We connect by edges each $x_{i,j}$ vertex to the $p_{i,j}$ vertex above it and to its 4 neighbors $x_{i \pm 1, j}, x_{i, j \pm 1}$ in the x-grid (except when the neighbor is off the grid) and no others. The cliques are just the pairs of vertices connected by edges. Finally, we

take for energies:

$$E_C = -x_{i,j} \cdot x_{i',j'}, \text{ when } C = \{(i,j),(i',j')\}, \text{ two adjacent vertices in the } x\text{-grid,}$$
$$E_C = -x_{i,j} \cdot y_{i,j}, \text{ when } C \text{ consists of the } (i,j) \text{ vertices in the } x\text{- and } y\text{-grids.}$$

The modes of the posteriors $\Pr_T(x_\cdot \mid \widehat{y}_\cdot)$ are quite subtle: x's at adjacent vertices try to be equal but they also seek to have the same sign as the correponding $\widehat{y}$. If $\widehat{y}$ has rapid positive and negative swings, these are in conflict. Hence the more probable values of x will align with the larger areas where $\widehat{y}$ is consistently of one sign. This can be used to model a basic problem in vision: the *segmentation problem.* The vision problem is to decompose the domain of an image y into parts where distinct objects are seen. For example, the oldman image might be decomposed into 6 parts: his body, his head, his cap, the bench, the wall behind him and the sky. The decomposition is to be based on the idea that the image will tend to either slowly varying or to be statistically stationary at points on one object, but to change abruptly at the edges of objects. As proposed in [G-G], the Ising model can be used to treat the case where the image has 2 parts, one lighter and one darker, so that at the mode of the posterior the hidden variables x will be $+1$ on one part, -1 on the other. An example is shown in figure 4. This approach makes a beautiful link between statistical mechanics and perception, in which the process of finding global patterns in a signal is like forming large scale structures in a physical material as the temperature cools through a phase transition.

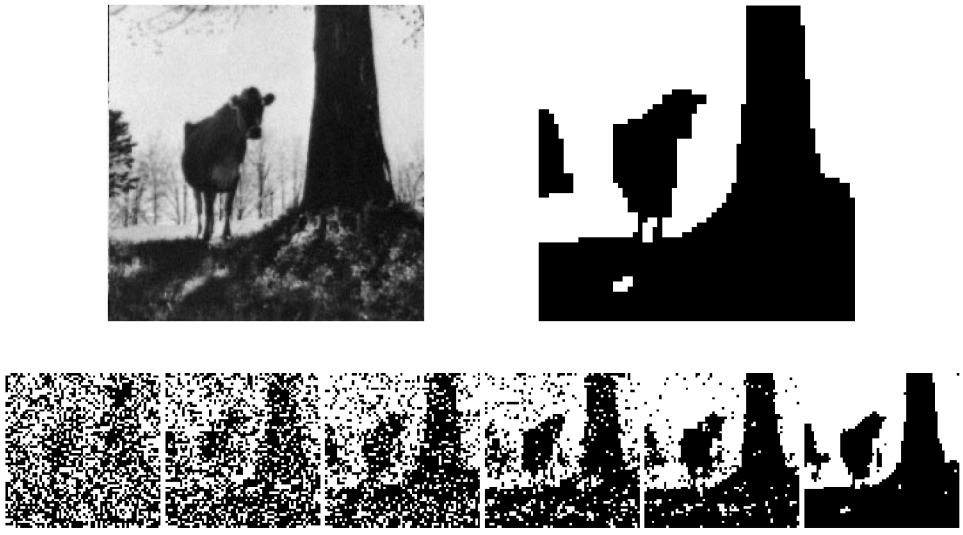


Figure 4: Statistical mechanics can be applied to the segmentation of images. On the top left, a rural scene taken as the external magnetic field, with its intensity scaled so that dark areas are negative, light areas are positive. At the top right, the mode or ground state of the Ising model. Along the bottom, the Gibbs distribution is sampled at a decreasing sequence of temperatures, discovering the global pattern bit by bit.

More complex models of this sort have been used extensively in image analysis, for texture segmentation, for finding disparity in stereo vision, for finding optic flow in moving images and for finding other kinds of groupings. We want to give one example of the expressivity of these models which is quite instructive. We saw above that exponential models can be crafted to reproduce some set of observed expectations but we also saw that scalar statistics from natural signals typically have high kurtosis, i.e. significant outliers, so that their whole distribution and not just their mean needs to be captured in the model. Putting these 2 facts together suggests that we seek exponential models which duplicate the whole distribution of some important statistics $f_\bullet$. This can be done using as parameters not just unknown constants but unknown functions:

$$\Pr(x_\bullet \mid \phi_\bullet) = \frac{1}{Z(\vec{\phi}_\bullet)} e^{\sum_k \phi_k(f_k(x_\bullet))}.$$

If f_k depends only the variables $x_v \in C_k$, for some clique C_k, this is a MRF, whose energies have unknown functions in them. An example of this fitting is shown in Figure 5.

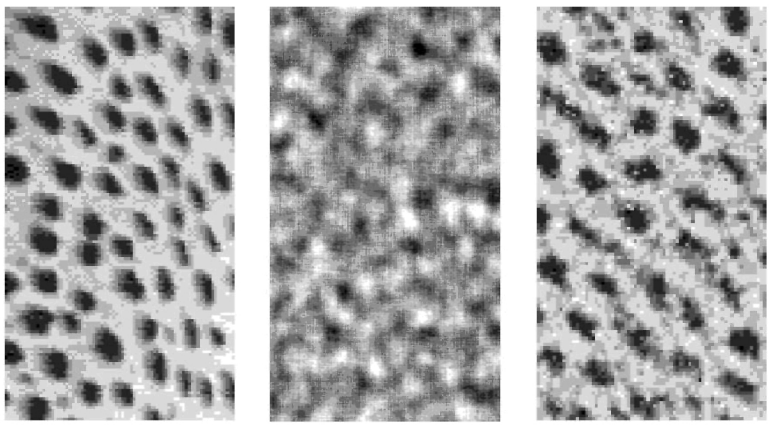


Figure 5: On the left, an image of the texture of a Cheetah's hide, in the middle a synthetic image from the Gaussian model with the same second order statistics, on the right a synthetic image in which the full distribution on 7 filter statistics are reproduced by an exponential model.

3.2. Bayesian belief propagation

However, a problem with MRF models is that the dynamic programming style algorithm used in speech and one-dimensional models to find the posterior mode has

no analog in 2D. One strategy for dealing with this, which goes back to Metropolis, is to imitate physics and introduce an artifical dynamics into the state space whose equilibrium is the Gibbs distribution. This dynamics is called a *Monte Carlo Markov Chain* (MCMC) and is how the panels in figure 4 were generated. Letting the temperature converge to zero, we get *simulated annealing* (see [G-G]) and, if we do it slowly enough, will find the mode of the MRF model. Although slow, this can be speeded up by biasing the dynamics (called *importance sampling* — see [T-Z] for a state-of-the-art implementation with many improvements) and is an important tool.

Recently, however, another idea due to Weiss and collaborators (see [Y-F-W]) and linked to statistical mechanics has been found to give new and remarkably effective algorithms for finding these modes. From an algorithmic standpoint, the idea is to use the natural generalization of dynamic programming, called *Bayesian Belief Propagation* (BBP), which computes the marginals and modes correctly whenever the graph is a tree and just use it anyway on an arbitrary graph G! Mathematically, it amounts to working on the universal covering graph $\widetilde{G}$, which is a tree, hence much simpler, instead of G. In statistical mechanics, this idea is called the *Bethe approximation*, introduced by him in the 30's.

To explain the idea, start with the *mean field approximation.* The mean field idea is to find the best approximation of the MRF p by a probability distribution in which the variables x_v are all independent. This is formulated as the distribution $\prod_v p_v(x_v)$ which minimizes the Kullback-Liebler divergence $\mathrm{KL}(\prod_v p_v, p)$. Unlike computing the true marginals of p on each x_v which is very hard, this approximation can be found by solving iteratively a coupled set of non-linear equations for the p_v. But the assumption of independence is much too restrictive. The idea of Bethe is instead to approximate p by a $\pi_1(G)$-invariant distribution on $\widetilde{G}$.

Such distributions are easy to describe: note that a Markov random field on a *tree* is uniquely determined by its marginals $p_e(x_v, x_w)$ for each edge $e = (v, w)$ and, conversely, if we are given a compatible set of distributions p_e for each edge (in the sense that, for all edges (v, w_k) abutting a vertex v, the marginals of $p_{(v,w_k)}$ give distributions on v independent of k), they define an MRF on G. So if we start with a Markov random field on any G, we get a $\pi_1(G)$-invariant Markov random field on $\widetilde{G}$ by making duplicate copies for each random variable $x_v, v \in V$ for each $\widetilde{v} \in \widetilde{V}$ over v and lifting the edge marginals. But more generally, if we have any compatible set of probability distributions $\{p_e(v, w)\}_{e \in E}$ on G, we also get a $\pi_1(G)$-invariant MRF on $\widetilde{G}$. Then the Bethe approximation is that family $\{p_e\}$ which minimizes $\mathrm{KL}(\{p_e\}, p)$. As in the mean field case, there is a natural iterative method of solving for this minimum, which turns out, remarkably, to be identical to the generalization of BBP to general graphs G.

This approach has proved effective in some cases at finding best segmentations of images via the mode of a two-dimensional MRF. Other interesting ideas have been proposed for solving the segmentation problem which we do not have time to sketch: region growing, see esp. [Z-Y]), using the eigenfunctions of the graph-theoretic Laplacian, see [S-M], and multi-scale algorithms, see [P-B] and [S-B-B].

4. Continuous space and time and continuous sets of random variables

Although signals as we measure them are always sampled discretely, in the world itself signals are functions on the continua, time or space or both together. In some situations, a much richer mathematical theory emerges by replacing a countable collection of random variables by random processes and asking whether we can find good stochastic models for these continuous signals. I want to conclude this talk by mentioning three instances where some interesting analysis has arisen when passing to the continuum limit and going into some detail on two. We will not worry about algorithmic issues for these models.

4.1. Deblurring and denoising of images

This is the area where the most work has been done, both because of its links with other areas of analysis and because it is one of the central problems of image processing. You observe a degraded image $I(x, y)$ as a function of continuous variables and seek to restore it, removing simultaneously noise and blur. In the discrete setting, the Ising model or variants thereof discussed above can be applied for this. There are two closely related ways to pass to the continuous limit and reformulate this as a problem in analysis. As both drop the stochastic interpretation and have excellent treatments in the literature, we only mention briefly one of a family of variants of each approach:

Optimal piecewise smooth approximation of I via a variational problem:

$$\min_{J,\Gamma} \left(c_1 \iint_D (I-J)^2 dxdy + c_2 \iint_{D-\Gamma} \|\nabla J\|^2 dxdy + c_3 |\Gamma| \right)$$

where J, the improved image, has discontinuities along the set of 'edge' curves Γ. This approach is due to the author and Shah and has been extensively pursued by the schools of DeGiorgi and Morel. See [M-S]. It is remarkable that it is still unknown whether the minima to this functional are well behaved, e.g. whether Γ has a finite number of components. Stochastic variants of this approach should exist.

Non-linear diffusion of I:

$$\frac{\partial J}{\partial t} = \operatorname{div}\left(\frac{\nabla J}{\|\nabla J\|} \right) + \lambda(I-J)$$

where J at some future time is the enhancement. This approach started with the work of Perona and Malik and has been extensively pursued by Osher and his coworkers. See [Gu-M]. It can be interpreted as gradient descent for a variant of the previous variational problem.

4.2. Self-similarity of image statistics and image models

One of the earliest discoveries about the statistics of images I was that their power spectra tend to obey power laws

$$\text{Exp}|(\widehat{I}(\xi,\eta)|^2 \approx (\xi^2+\eta^2)^{-\lambda/2},$$

where λ varies somewhat from image to image but clusters around the value 2. This has a very provocative interpretation: this power law is implied by self-similarity! In the language of lattice field theory, if $I(i,j), i,j \in \mathbb{Z}$ is a random lattice field and $\bar{I}$ is the block averaged field

$$\bar{I}(i,j) = \frac{1}{4}\left(I(2i,2j)+I(2i+1,2j)+I(2i,2j+1)+I(2i+1,2j+1)\right),$$

then we say the field is a renormalization fixed point if the distributions of I and of $\bar{I}$ are the same. The hypothesis that natural images of the world, treated as a single large database, have renormalization invariant statistics has received remarkable confirmation from many quite distinct tests.

Why does this hold? It certainly isn't true for auditory or tactile signals. I think there is one major and one minor reason for it. The major one is that the world is viewed from a random viewpoint, so one can move closer or farther from any scene. To first approximation, this scales the image (though not exactly because nearer objects scale faster than distant ones). The minor one is that most objects are opaque but have, by and large, parts or patterns on them and, in turn, belong to clusters of larger things. This observation may be formulated as saying the world is not merely made up of objects but it is cluttered with them.

The natural setting for scale invariance is pass to the limit and model images as random functions $I(x,y)$ of two real variables. Then the hypothesis is that a suitable function space supports a probability measure which is invariant under both translations and scalings $(x,y) \mapsto (\sigma x, \sigma y)$, whose samples are 'natural images'. This hypothesis encounters, however, an infra-red and an ultra-violet catastrophe:
a) The infra-red one is caused by larger and larger scale effects giving bigger and bigger positive and negative swings to a local value of I. But these large scale effects are very low-frequency and this is solved by considering I to be defined only modulo an unknown constant, i.e. it is a sample from a measure on a function space mod constants.
b) The ultra-violet one is worse: there are more and more local oscillations of the signals at finer and finer scales and this contradicts Lusin's theorem that an integrable function is continuous outside sets of arbitrarily small measure. In fact, it is a theorem that *there is no translation and scale invariant probability measure on the space of locally integrable functions mod constants.* This can be avoided by allowing images to be generalized functions. In fact, the support can be as small as the intersection of all negative Sobolev spaces $\bigcap_\epsilon \mathcal{H}^{-\epsilon}$.

To summarize what a good statistical theory of natural images should explain, we have scale-invariance as just described, kurtosis greater than 3 as described in section 2.1 and finally the right local properties:

Hypothesis I A theory of images is a translation and scale invariant probability measure on the space of generalized functions $I(x, y)$ mod constants.

Hypothesis II For any filter F with mean 0, the marginal statistics of $F * I(x, y)$ have kurtosis greater than 3.

Hypothesis III The local statistics of images reflect the preferred local geometries, esp. images of straight edges, but also curved edges, corners, bars, 'T-junctions' and 'blobs' as well as images without geometry, blank 'blue sky' patches.

Hypothesis III is roughly the existence of what Marr, thinking globally of the image called the *primal sketch* and what Julesz, thinking locally of the elements of texture, referred to as *textons.* By scale invariance, the local and global image should have the same elements.

To quantify Hypothesis III, what is needed is a major effort at data mining. Specifically, the natural approach seems to be to take a small filter bank of zero mean local filters $F_1, \cdots, F_k$, a large data base of natural images I_α leading to the sample of points in $\mathbb{R}^k$ given by $(F_1 * I_\alpha(x, y), \cdots, F_K * I_\alpha(x, y)) \in \mathbb{R}^k$ for all α, x and y. One seeks a good non-parametric fit to this dataset. But Hypothesis III shows that this distribution will not be simple. For example Lee et al [L-P-M] have taken $k = 8$, F_i a basis of zero mean filters with fixed 3×3 support. They then make a linear tranformation in $\mathbb{R}^8$ normalizing the covariance of the data to I_8 ('whitening' the data), and to investigate the outliers, map the data with norms in the upper 20% to S^7 by dividing by the norm. The analysis reveals that the resulting data has asymptotic infinite density along a non-linear surface in S^7! This surface is constructed by starting with an ideal image, black and white on the two sides of a straight edge and forming a 3×3 discrete image patch by integrating this ideal image over a tic-tac-toe board of square pixels. As the angle of the edge and the offset of the pixels to the edge vary, the resulting patches form this surface. This is the most concrete piece of evidence showing the complexity of local image statistics.

Are there models for these three hypotheses? We can satisfy the first hypothesis by the unique scale-invariant Gaussian model, called the free field by physicists — but its samples look like clouds and its marginals have kurtosis 3, so neither the second nor third hypothesis is satisfied. The next best approximation seems to be to use infinitely divisible measures, such as the model constructed by the author and B.Gidas [M-G], which we call *random wavelet expansions*:

$$I(x, y) = \sum_i \phi_i(e^{r_i} x - x_i, e^{r_i} y - y_i),$$

where $\{(x_i, y_i, r_i)\}$ is a Poisson process in $\mathbb{R}^3$ and ϕ_i are samples from an auxiliary Levi measure, playing the role of individual random wavelet primitives. But this model is based on adding primitives, as in a world of transparent objects, which causes the probability density functions of its marginal filter statistics to be smooth at 0 instead of having peaks there, i.e. the model does not produce enough 'blue sky' patches with very low constrast.

A better approach are the random collage models, called *dead leaves models* by the French school: see [L-M-H]. Here the ϕ_i are assumed to have bounded support, the terms have a random depth and, instead of being simply added, each term occludes anything behind it with respect to depth. This means $I(x,y)$ equals the one ϕ_i which is in front of all the others whose support contains (x,y). This theory has major troubles with both infra-red and ultra-violet limits but it does provide the best approximation to date of the empirical statistics of images. It introduces explicitly the hidden variables describing the discrete objects in the image and allows one to model their preferred geometries.

Crafting models of this type is not simply mathematically satisfying. It is central to the main application of computer vision: object recognition. When an object of interest is obscured in a cluttered badly lit scene, one needs a p-value for the hypothesis test — is this fragment of stuff part of the sought-for object or an accidental conjunction of things occurring in generic images? To get this p-value, one needs a null hypothesis, a theory of generic images.

4.3. Stochastic shapes via random diffeomorphisms and fluid flow

As we have seen in the last section, modeling images leads to objects and these objects have shape — so we need stochastic models of shape, the ultimate non-linear sort of thing. Again it is natural to consider this in the continuum limit and consider a k-dimensional shape to be a subset of $\mathbb{R}^k$, e.g. a connected open subset with nice boundary Γ. It is very common in multiple images of objects like faces, animals, clothes, organs in your body, to find not identical shapes but warped versions. How is this to be modeled? One can follow the ideas of the previous section and take a highly empirical approach, gathering huge databases of faces or kidneys. This is probably the road to the best pattern recognition in the long run. But another principle that Grenander has always emphasized is to take advantage of the group of symmetries of the situation — in this case, the group of all diffeomorphisms of $\mathbb{R}^k$. He and Miller and collaborators (see [Gr-M]) were led to rediscover the point of view of Arnold which we next describe.

Let $\mathcal{G}_n =$ group of diffeomorphisms on $\mathbb{R}^n$ and $\mathcal{SG}_n$ be the volume-preserving subgroup. We want to bypass issues of the exact degree of differentiability of these diffeomorphisms, but consider $\mathcal{G}_n$ and $\mathcal{SG}_n$ as infinite dimensional Riemannian manifolds. Let $\{\theta_t\}_{0\le t\le 1}$ be a path in $\mathcal{SG}_n$ and define its length by:

$$\text{length of path } = \int \left(\sqrt{\int_{\mathbb{R}^n} \| \frac{\partial \theta_t}{\partial t}(\theta_t^{-1}(x)) \|^2 d\vec{x}} \right) dt.$$

This length is nothing but the *right*-invariant Riemannian metric:

$$\text{dist}(\theta, (I + \epsilon\vec{v}) \circ \theta)^2 = \epsilon^2 \int \|\vec{v}\|^2 dx_1 \cdot \cdot dx_n, \text{ where } \text{div}(\vec{v}) \equiv 0.$$

Arnold's beautiful theorem is:

Theorem *Geodesics in $\mathcal{SG}_n$ are solutions of Euler's equation:*

$$\frac{\partial v_t}{\partial t} + (v_t \cdot \nabla)v_t = \nabla p, \text{ some pressure } p.$$

This result suggests using geodesics on suitable infinite dimensional manifolds to model optimal warps between similar shapes in images and using diffusion on these manifolds to craft stochastic models. But we need to get rid of the volume-preserving restriction. The weak metric used by Arnold no longer works on the full $\mathcal{G}_n$ and in [C-R-M], Christensen et al introduced:

$$\|\vec{v}\|_L^2 = \int < L\vec{v} \cdot \vec{v} > dx_1 \cdots dx_n$$

where v is any vector field and L is a fixed positive self-adjoint differential operator e.g. $(I - \Delta)^m, m > n/2$. Then a path $\{\theta_t\}$ in G has both a velocity:

$$v_t = \frac{\partial \theta_t}{\partial t}(\theta_t^{-1}(x))$$

and a *momentum:* $u_t = Lv_t$ (so $v_t = K * u_t$, K the Green's function of L). What is important here is that the momentum u_t can be a generalized function, even when v_t is smooth. The generalization of Arnold's theorem, first derived by Vishik, states that geodesics are:

$$\frac{\partial u_t}{\partial t} + (v_t \cdot \nabla)(u_t) + \text{div}(v_t)u_t = -\sum_i (u_t)_i \vec{\nabla}((v_t)_i).$$

This equation is a new kind of regularized compressible Euler equation, called by Marsden the template matching equation (TME). The left hand side is the derivative along the flow of the momentum, as a measure, and the right hand side is the force term.

A wonderful fact about this equation is that by making the momentum singular, we get very nice equations for geodesics on the $\mathcal{G}_n$-homogeneous spaces:

(a) $\mathcal{L}_n$ = set of all N-tuples of distinct points in $\mathbb{R}^n$ and

(b) $\mathcal{S}_n$ = set of all images of the unit ball under a diffeomorphism.

In the first case, we have $\mathcal{L}_n \cong \mathcal{G}_n/\mathcal{G}_{n,0}$ where $\mathcal{G}_{n,0}$ is the stabilizer of a specific set $\{P_1^{(0)}, \cdots, P_N^{(0)}\}$ of N distinct points. To get geodesics on $\mathcal{L}_n$, we look for 'particle solutions of the TME', i.e.

$$\vec{u}_t = \sum_{i=1}^{N} \vec{u}_i(t)\delta_{P_i(t)}$$

where $\{P_1(t), \cdots, P_N(t)\}$ is a path in $\mathcal{L}_n$ The geodesics on $\mathcal{G}_n$, which are perpendicular to all cosets $\theta\mathcal{G}_{n,0}$, are then the geodesics on $\mathcal{L}_n$ for the quotient metric:

$$\begin{aligned} \text{dist}(\{P_i\}, \{P_i + \epsilon v_i\})^2 &= \epsilon^2 \inf_{\substack{v \text{ on } \mathbb{R}^n \\ v(P_i)=v_i}} \int < Lv, v > \\ &= \epsilon^2 \sum_{i,j} G_{ij}(v_i \cdot v_j) \end{aligned}$$

where $G = K(\|P_i - P_j\|)^{-1}$. For these we get the interesting Hamiltonian ODE:

$$\begin{aligned} \frac{dP_i}{dt} &= 2\sum_j K(\|P_i - P_j\|)\vec{u}_j \\ \frac{du_i}{dt} &= -\sum_j \nabla_{P_i} K(\|P_i - P_j\|) \cdot (\vec{u}_i \cdot \vec{u}_j) \end{aligned}$$

which makes points traveling in the same direction attract each other and points going in opposite directions repel each other. This space leads to a non-linear version of the theory of landmark points and shape statistics of Kendall [Sm] and has been developed by Younes [Yo].

A similar treatment can be made for the space of shapes $\mathcal{S}_n \cong \mathcal{G}_n/\mathcal{G}_{n,1}$, where $\mathcal{G}_{n,1}$ is the stabilizer of the unit sphere. Geodesics on $\mathcal{S}_n$ come from solutions of the TME for which $\vec{u}_t$ is supported on the boundary of the shape and perpendicular to it. Even though the first of these spaces $\mathcal{S}_2$ might seem to be quite a simple space, it seems to have a remarkable global geometry reflecting the many perceptual distinctions which we make when we recognize a similar shapes, e.g. a cell decomposition reflecting the different possible graphs which can occur as the 'medial axis' of the shape. This is an area in which I anticipate interesting results. We can also use these Riemannian structures to define Brownian motion on $\mathcal{G}_n, \mathcal{S}_n$ and $\mathcal{L}_n$ (see [D-G-M], [Yi]). Putting a random stopping time on this walk, we get probability measures on these spaces. To make the ideas more concrete, in figure 6 we show a simulation of the random walk on $\mathcal{S}_2$.

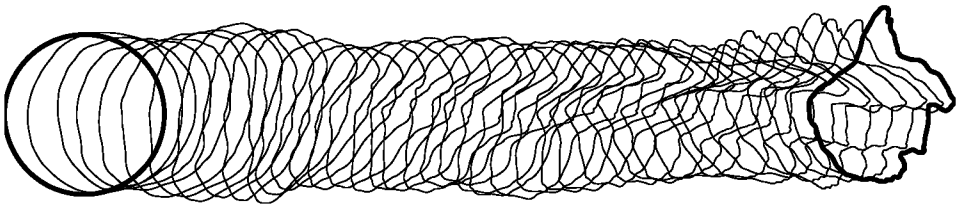


Figure 6: An example of a random walk in the space of 2D shapes $\mathcal{S}_2$. The initial point is the circle on the left. A constant translation to the right has been added so the figures can be distinguished. The operator L defining the metric is $(I - \triangle)^2$

5. Final thoughts

The patterns which occur in nature's sensory signals are complex but allow mathematical modeling. Their study has gone through several phases. At first, 'off-the-shelf' classical models (e.g. linear Gaussian models) were adopted based only on intuition about the variability of the signals. Now, however, two things are happening: computers are large enough to allow massive data gathering to support fully non-parametric models. And the issues raised by these models are driving the study of new areas of mathematics and the development of new algorithms for working with these models. Applications like general purpose speech recognizers and

computer driven vehicles are likely in the foreseeable future. Perhaps the ultimate dream is a fully unsupervised learning machine which is given only signals from the world and which finds their statistically significant patterns with no assistance: something like a baby in its first 6 months.

References

[B-G-P] E Bienenstock, S Geman and D Potter, Compositionality, MDL priors and object recognition, *Adv. in Neural Information Proc.*, Mozer, Jordan and Petsche ed., **9**, MIT Press, 1998.

[C-R-M] G Christensen, RD Rabbitt and M Miller, 3D brain mapping using a deformable neuroanatomy, *Physics in Med. and Biol.*, **39**, 1994.

[D-F-G] A Doucet, N. de Freitas and N Gordon editors, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[D-G-M] P Dupuis, U Grenander and M Miller, Variational problems on flows of diffeomorphisms for image matching, *Quarterly Appl. Math.*, **56**, 1998.

[G] U Grenander, *General Pattern Theory*, Oxford Univ. Press, 1993.

[G-G] S Geman and D Geman, Stochastic relaxation, Gibbs distr. and Bayesian restoration of images, **PAMI** = *IEEE Trans. Patt. Anal. Mach. Int.*, **6**, 1984.

[Gr-M] U Grenander and M Miller, Computational anatomy, *Quart. of Appl. Math.*, **56**, 1998.

[Gu-M] F Guichard and J-M Morel, *Image Anal. and Partial Diff. Equations*, `http://www.ipam.ucla.edu/publications/gbm2001/gbmtut_jmorel.pdf`.

[G-S-S] N Gordon, D Salmond and A Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proc.-F*, **140**, 1993.

[H] L.C.G. Haggerty, What and two-and-a-half year old child said in one day, *J. Genet. Psych.*, **37**, 1930.

[I-B] M Isard and A Blake, Contour tracking by stochastic propagation of conditional density, *Eur. Conf. Comp. Vis.*, 1996.

[L] P Lieberman, *The Biology and Evolution of Language*, Harvard, 1984.

[L-M-H] A Lee, D Mumford and J Huang, An occlusion model for natural images, *Int. J. Comp. Vis.*, **41**, 2001.

[L-P-M] A Lee, K Pederson and D Mumford, The non-linear statistics of high-contrast patches in natural images, to appear, *Int. J. Comp. Vis.*.

[M] D Mumford, The Dawning of the Age of Stochasticity, in *Mathematics, Frontiers and Perspectives*, ed. by Arnold, Atiyah, Lax and Mazur, AMS, 2000.

[M-G] D Mumford and B Gidas, Stochastic models for generic images, *Quarterly of Appl. Math.* , **59**, 2001.

[M-S] J-M Morel and S Solimini, *Variational Methods in Image Segmentation*, Birkhauser, 1995.

[P] J Pearl, *Probabilistic Reasoning in Int. Systems*, Morgan-Kaufmann, 1997.

[P-B] J Puzicha and J Buhmann, Multiscale annealing for grouping and texture

segmentation, *Comp. Vis. and Image Understanding*, **76**, 1999.

[Sh] S Shieber, *Constraint-Based Grammar Formalisms*, MIT Press, 1992.

[Sm] C Small, *The Statistical Theory of Shape*, Springer, 1996.

[S-B-B] E Sharon, A Brandt and R Basri, Segmentation and boundary detection using multiscale intensity measurements, *Proc IEEE Conf. Comp. Vis. Patt. Recognition*, Hawaii, 2001.

[S-M] J Shi and J Malik, Normalized cuts and image segmentation, *PAMI*, **22**, 2000.

[T-Z] -W Tu and S-C Zhu, Image segmentation by data driven Markov chain Monte Carlo, *PAMI*, **24**, 2002.

[V] V Vapnik, *The Nature of Statistical Learning Theory*, Springer, 199?.

[W] R.M. Warren, Restoration of missing speech sounds, *Science*, **167**, 1970.

[Yi] N-K Yip, Stoch. motion by mean curv., *Arch. Rat. Mech. Anal.*, **144**, 1998.

[Yo] L Younes, *Invariance, Déformations et Reconnaissance de Formes*, to appear.

[Y-F-W] J Yedidia, W Freeman, and Y Weiss, Generalized Belief Propagation, *Adv. in Neural Information Proc.*, edited by Leen, Dietterich, Tresp, **13**, 2001.

[Z-Y] S-C Zhu and A Yuille, Region competition, *PAMI*, **18**, 1996.

ICM 2002 · Vol. I · 423–438

Geometric Construction of Representations of Affine Algebras

Hiraku Nakajima*

Abstract

Let Γ be a finite subgroup of $\mathrm{SL}_2(\mathbf{C})$. We consider Γ-fixed point sets in Hilbert schemes of points on the affine plane $\mathbf{C}^2$. The direct sum of homology groups of components has a structure of a representation of the affine Lie algebra $\widehat{\mathfrak{g}}$ corresponding to Γ. If we replace homology groups by equivariant K-homology groups, we get a representation of the quantum toroidal algebra $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$. We also discuss a higher rank generalization and character formulas in terms of intersection homology groups.

2000 Mathematics Subject Classification: 17B37, 14D21, 14L30, 16G20, 33D80.
Keywords and Phrases: Affine Lie algebras, Quantum toroidal algebras, Hilbert schemes, Quiver varieties.

1. Finite subgroups of $\mathrm{SL}_2(\mathrm{C})$ and simple Lie algebras

Let Γ be a finite subgroup of $\mathrm{SL}_2(\mathbf{C})$. The classification of such subgroups has been well-known to us, since they are essentially symmetry groups of regular polytopes. They are cyclic groups, binary dihedral groups, and binary polyhedral groups (Klein (1884)).

It has been also known that we can associate a complex simple Lie algebra $\mathfrak{g}$ to Γ. This can be done in two ways. The first one is geometric and due to DuVal (1934). The second one is algebraic, and is due to McKay (1979).

Let us explain the two constructions and subsequent developments briefly. More detailed account can be found in [17].

1.1. Minimal resolution of $\mathbf{C}^2/\Gamma$

*Department of Mathematics, Kyoto University, Kyoto 606-8502, Japan. E-mail: nakajima@kusm.kyoto-u.ac.jp

Let us consider the quotient space $\mathbf{C}^2/\Gamma$. This space has an isolated singularity at the origin. We have a unique *minimal* resolution $\pi\colon M \to \mathbf{C}^2/\Gamma$, in the sense that all other resolutions factor through π. (For general singularities, we have many resolutions. This speciality occurs in 2-dimensional case.) This singularity is called a *simple singularity*, and has been intensively studied from various points of view. In particular, the following are known (see e.g., [2]):

1. The exceptional set $\pi^{-1}(0)$ consists of the union of projective lines.
2. We draw a diagram so that vertices correspond to projective lines (irreducible components) and two vertices are connected by an edge if they intersect. Then we obtain a Dynkin diagram of type ADE.

We thus have bijections

$$\{\text{irreducible components of } \pi^{-1}(0)\} \longleftrightarrow \{\text{vertices of the Dynkin diagram}\}.$$

The Dynkin diagram appears in the classification of simple Lie algebras. Thus we have a complex simple Lie algebra $\mathfrak{g}$ corresponding to Γ. Since vertices of the Dynkin diagram correspond to simple coroots of $\mathfrak{g}$, the above bijection gives an isomorphism (of vector spaces)

$$\mathfrak{h} \xrightarrow{\sim} H_2(\pi^{-1}(0), \mathbf{C}), \tag{1.1}$$

where $\mathfrak{h}$ is the complex Cartan subalgebra of $\mathfrak{g}$.

This correspondence $\Gamma \to \mathfrak{g}$ is based on the classification of simple Lie algebras since they attach a Dynkin diagram to Γ. So the reason why such a result holds remained misterious. A deeper connection between two objects were conjectured by Grothendieck, and obtained by Brieskorn (1970) and Slodowy (1980). They constructed the simple singularity $\mathbf{C}^2/\Gamma$ in $\mathfrak{g}$. Moreover, its semi-universal deformation and a simultaneous resolution were also constructed using geometry related to $\mathfrak{g}$. We do not recall their results here, so the interested reader should consult [40].

1.2. McKay correspondence

Let $\{\rho_i\}_{i\in I}$ be the set of (isomorphism classes of) irreducible representations of Γ. It has a special element ρ_0, the class of trivial representation. Let Q be the 2-dimensional representation given by the inclusion $\Gamma \subset \mathrm{SL}_2(\mathbf{C})$. Let us decompose $Q \otimes \rho_i$ into irreducibles, $Q \otimes \rho_i = \bigoplus_j a_{ij}\rho_j$, where a_{ij} is the multiplicity. We draw a diagram so that vertices correspond to ρ_i's, and there are a_{ij} edges between ρ_i and ρ_j. (Note that $a_{ij} = a_{ji}$ thanks to the self-duality of Q). Then McKay [26] observed that the graph is an affine Dynkin diagram of $\tilde{A}_n^{(1)}, \tilde{D}_n^{(1)}, \tilde{E}_6^{(1)}, \tilde{E}_7^{(1)}$ or $\tilde{E}_8^{(1)}$, i.e., the Dynkin diagram of an untwisted affine Lie algebra $\widehat{\mathfrak{g}}$ attached to a simple Lie algebra $\mathfrak{g}$ of type ADE. Furthermore it is also known that the Dynkin diagram given in the previous subsection is obtained by the affine Dynkin diagram by removing the vertex corresponding to the trivial representation ρ_0. We thus have bijections

$$\{\text{irreducible representations of } \Gamma\} \longleftrightarrow \{\text{vertices of the affine Dynkin diagram}\}.$$

The original McKay's proof was based on the explicit calculation of characters. The reason why such a result holds remained misterious also in this case. A geometric explanation via the K-theory of the minimal resolution M of $\mathbf{C}^2/\Gamma$ was subsequently given by Gonzalez-Sprinberg and Verdier [14]. In particular, they proved that there exists a natural geometric construction of an isomorphism (of abelian groups)

$$R(\Gamma) \xrightarrow{\sim} K(M),$$

where $R(\Gamma)$ is the representation ring of Γ, and $K(M)$ is the Grothendieck group of the abelian category of algebraic vector bundles over M. This result is strengthened and generalized to the higher dimensional case $\Gamma \subset \mathrm{SL}_3(\mathbf{C})$ [5].

Note that the above isomorphism together with the Chern character homomorphism leads to an isomorphism $R(\Gamma) \otimes_{\mathbf{Z}} \mathbf{C} \xrightarrow{\sim} H^*(M, \mathbf{C})$, which gives an isomorphism

$$R(\Gamma) \otimes_{\mathbf{Z}} \mathbf{C} \xrightarrow{\sim} (\mathfrak{h} \oplus \mathbf{C}h_0)^*, \tag{1.2}$$

combined with (1.1). Here h_0 is the 0th simple coroot of the affine Lie algebra $\widehat{\mathfrak{g}}$, and corresponds to the dual of the trivial representation ρ_0. It corresponds to $H_0(M, \mathbf{C}) \cong H_0(\pi^{-1}(0), \mathbf{C})$.

Compared with correspondence in §1.1, our situation is less satisfactory: we only get $\mathfrak{h}$ and the role of $\mathfrak{g}$ or $\widehat{\mathfrak{g}}$ is less clear. This is the starting point of our whole construction. We construct $\widehat{\mathfrak{g}}$ entirely from Γ in some sense. For another approach, see [13].

2. Hilbert schemes of points and their Γ-fixed point components — quiver varieties

In 1986, Kronheimer [20] constructed a simple singularity $\mathbf{C}^2/\Gamma$, its deformation and simultaneous resolution, i.e., those spaces constructed by Brieskorn-Slodowy by a totally different method. His construction is based on the theory of 'quivers', which is a subject in noncommutative algebras. (See also [6] for a different approach.) Subsequently in 1989, Kronheimer and the author [21] gave a description of moduli spaces of instantons (and coherent sheaves) on those spaces in terms of a quiver. It is an analog of the celebrated ADHM description of instantons on S^4. In 1994, this description was further generalized under the name of 'quiver varieties' by the author [27]. The purpose of this and next sections is to define quiver varieties from a slightly different point of view. This is a most economical approach to introduce quiver varieties, while it does not explain why it is something to do with quivers.

Let $\mathrm{Hilb}^n(\mathbf{C}^2)$ be the Hilbert scheme of n points in the affine plane $\mathbf{C}^2$. As a set, it consists of ideals I of the polynomial ring $\mathbf{C}[x, y]$ such that the quotient $\mathbf{C}[x, y]/I$ has dimension n as a vector space. Grothendieck constructed $\mathrm{Hilb}^n(\mathbf{C}^2)$ as a quasi-projective scheme (for more general setting), but we do not go to this direction in detail. A typical point of $\mathrm{Hilb}^n(\mathbf{C}^2)$ is an ideal of functions vanishing at n distinct points in $\mathbf{C}^2$. The space parametrizing (unordered) n distinct points is an open subset of the nth symmetric product $S^n(\mathbf{C}^2) = (\mathbf{C}^2)^n/S_n$ of $\mathbf{C}^2$, where

S_n is the symmetric group of n letters acting on $(\mathbf{C}^2)^n$ by permutation of factors. The symmetric product parametrises unordered n points with multiplicities. The Hilbert scheme $\mathrm{Hilb}^n(\mathbf{C}^2)$ is a different completion of the open set. Two completions are related: Mapping I to its support counted with multiplicities, we get a morphism $\pi\colon \mathrm{Hilb}^n(\mathbf{C}^2) \to S^n(\mathbf{C}^2)$ which is called a *Hilbert-Chow morphism.* We have following important geometric results on $\mathrm{Hilb}^n(\mathbf{C}^2)$:

1. $\mathrm{Hilb}^n(\mathbf{C}^2)$ is a resolution of singularities of $S^n(\mathbf{C}^2)$ (Fogarty).
2. $\mathrm{Hilb}^n(\mathbf{C}^2)$ has a holomorphic symplectic structure (Beauville, Mukai).

In fact, the author constructed a hyper-Kähler structure on $\mathrm{Hilb}^n(\mathbf{C}^2)$, which induces Beauville-Mukai's symplectic form, by describing it as a hyper-Kähler quotient. See [29] and Göttsche's article in this ICM proceeding for more recent results on $\mathrm{Hilb}^n(\mathbf{C}^2)$.

Let Γ be a finite subgroup of $\mathrm{SL}_2(\mathbf{C})$ as above. Its natural action on $\mathbf{C}^2$ induces an action on $\mathrm{Hilb}^n(\mathbf{C}^2)$ and $S^n(\mathbf{C}^2)$ such that the Hilbert-Chow morphism π is Γ-equivariant. Let us consider the fixed point set $\mathrm{Hilb}^n(\mathbf{C}^2)^\Gamma$, $(S^n(\mathbf{C}^2))^\Gamma$. The latter is easy to describe:

$$(S^n(\mathbf{C}^2))^\Gamma = S^m(\mathbf{C}^2/\Gamma),$$

where m is the largest integer less than or equal to $n/\#\Gamma$. The difference $n - m\#\Gamma$ is the multiplicity of the origin. The former space $\mathrm{Hilb}^n(\mathbf{C}^2)^\Gamma$ is a union of nonsingular submanifolds of $\mathrm{Hilb}^n(\mathbf{C}^2)$. If $I \in \mathrm{Hilb}^n(\mathbf{C}^2)^\Gamma$, the quotient $\mathbf{C}[x,y]/I$ has a structure of a representation of Γ. For an isomorphism class $\mathbf{v}$ of a representation of Γ, we define $M(\mathbf{v})$ as

$$M(\mathbf{v}) = \left\{ I \in \mathrm{Hilb}^n(\mathbf{C}^2)^\Gamma \;\middle|\; [\mathbf{C}[x,y]/I] = \mathbf{v} \right\},$$

where $[\mathbf{C}[x,y]/I]$ is the isomorphism class of $\mathbf{C}[x,y]/I$. Since isomorphism classes are parametrized by discrete data, i.e., dimensions of isotropic components, the isomorphism class of $[\mathbf{C}[x,y]/I]$ is constant on each connected component. Therefore $M(\mathbf{v})$ is a union of connected component. In fact, Crawley-Boevey recently proves that $M(\mathbf{v})$ is connected (in fact, he proved it for more general case including varieties discussed in next section) [9]. Moreover, $M(\mathbf{v})$ has induced holomorphic symplectic and hyper-Kähelr structures. It is an example of quiver varities of affine type. (See remark at the end of the next section.)

The simplest but nontrivial example is the case when $\mathbf{v}$ is the class of the regular representation of Γ. Under (1.2), the regular representation corresponds to the imaginary root δ, which is the positive generator of the kernel of the affine Cartan matrix, is identified with the dimension vector of the regular representation of Γ. The dimension of the regular representation is equal to $\#\Gamma$, and thus the fixed point set in the symmetric product is $(S^{\#\Gamma}(\mathbf{C}^2))^\Gamma = \mathbf{C}^2/\Gamma$. We can consider this as the space of Γ-orbits. A typical point is a free Γ-orbit, and is also a point in $\mathrm{Hilb}^{\#\Gamma}(\mathbf{C}^2)^\Gamma$ as the ideal vanishing at the orbit. In fact, it is not difficult to see that $M(\mathbf{v})$ is isomorphic to the minimal resolution M of $\mathbf{C}^2/\Gamma$. The resolution map $\pi\colon M \to \mathbf{C}^2/\Gamma$ is given by the restriction of the Hilbert-Chow morphism. This

result was obtained by Ginzburg-Kapranov (unpublished) and Ito-Nakamura [17] independently, but is also a re-interpretation of Kronheimer's construction [20]. The precise explanation was given in [29, Chapter 4].

Recently higher dimensional $M(\mathbf{v})$ attract attention in connection with the McKay correspondence for wreath products $\Gamma \wr S_n$ [44, 22, 15]. These $M(\mathbf{v})$ are diffeomorphic to the Hilbert schemes of points on the minimal resolution $\mathrm{Hilb}^n M$.

3. A higher rank generalization of Hilbert schemes

We give a higher rank generalization of Hilbert schemes in this section. But geometric structures remain unchanged for general cases. So a reader, who wants to catch only a *rough* picture, could safely skip this section.

Let $\mathbf{P}^2$ be the projective plane with a fixed line ℓ_∞. So $\mathbf{P}^2 = \mathbf{C}^2 \sqcup \ell_\infty$. Let $\mathfrak{M}(n,r)$ be the framed moduli space of torsion free sheaves on $\mathbf{P}^2$ with rank r and $c_2 = n$, i.e. the set of isomorphism classes of pairs (E,φ), where E is a torsion free sheaf of $\operatorname{rank} E = r$, $c_2(E) = n$, which is locally free in a neighbourhood of ℓ_∞, and φ is an isomorphism $\varphi\colon E|_{\ell_\infty} \xrightarrow{\sim} \mathcal{O}_{\ell_\infty}^{\oplus r}$ (framing at infinity). It is known that this space has a structure of a quasi-projective variety [16]. This is a higher rank generalization of the Hilbert scheme $\mathrm{Hilb}^n(\mathbf{C}^2)$. The analog of $\mathbf{C}[x,y]/I$ is $H^1(\mathbf{P}^2, E(-1))$ and it is known that $H^0(\mathbf{P}^2, E(-1)) = H^2(\mathbf{P}^2, E(-1)) = 0$ [29, Chapter 2]. It is also known that $\mathfrak{M}(n,r)$ has a holomorphic symplectic (in fact, hyper-Kähler) structure [29, Chapter 3].

The higher rank generalization of the symmetric product $S^n(\mathbf{C}^2)$ is the so-called Uhlenbeck compactification of the framed moduli space of locally free sheaves. (On the other hand, $\mathfrak{M}(n,r)$ is called Gieseker-Maruyama compactification.) It is

$$\mathfrak{M}_0(n,r) = \bigsqcup_{n'+n''=n} \mathfrak{M}_0^{\mathrm{reg}}(n',r) \times S^{n''}\mathbf{C}^2,$$

where $\mathfrak{M}_0^{\mathrm{reg}}(n',r)$ is the open subset of $\mathfrak{M}(n',r)$ consisting of framed *locally free* sheaves (E,φ). It is known that $\mathfrak{M}_0(n,r)$ has a structure of an affine algebraic variety [10, Chapter 3]. Moreover, the map

$$(E,\varphi) \mapsto (E^{\vee\vee}, \varphi, \mathrm{Supp}(E^{\vee\vee}/E))$$

gives a projective morphism $\pi\colon \mathfrak{M}(n,r) \to \mathfrak{M}_0(n,r)$ [29, Chapter 3], where $E^{\vee\vee}$ is the double dual of E, which is locally free on surfaces, and $\mathrm{Supp}(E^{\vee\vee}/E)$ is the support of $E^{\vee\vee}/E$, counted with multiplicities.

When $r = 1$, there exists only one locally free sheaf which is trivial at ℓ_∞, i.e., the trivial line bundle $\mathcal{O}_{\mathbf{P}^2}$. So the first factor of the above disappears : $\mathfrak{M}_0(n,1) = S^n\mathbf{C}^2$. Moreover, for $E \in \mathfrak{M}(n,1)$, the double dual $E^{\vee\vee}$ must be the trivial line bundle by the same reason. It means that E is an ideal sheaf of the structure sheaf $\mathcal{O}_{\mathbf{P}^2}$, so is a point in the Hilbert scheme $\mathrm{Hilb}^n(\mathbf{C}^2)$. Thus we recover the situation studied in §2.

Let Γ be a finite subgroup of $\mathrm{SL}_2(\mathbf{C})$ as before. We take and fix a lift of the Γ-action to $\mathcal{O}_{\ell_\infty}^{\oplus r}$. It is written as $W \otimes_{\mathbf{C}} \mathcal{O}_{\ell_\infty}$, where W is a representation

W of Γ. We denote by $\mathbf{w}$ the isomorphism class of W as before. Now Γ acts on $\mathfrak{M}(n,r)$, $\mathfrak{M}_0(n,r)$ and we can consider the fixed point sets $\mathfrak{M}(n,r)^\Gamma$, $\mathfrak{M}_0(n,r)^\Gamma$. We decompose the former as

$$\mathfrak{M}(n,r)^\Gamma = \bigsqcup_{\mathbf{v}} \mathfrak{M}(\mathbf{v},\mathbf{w}),$$

where $\mathfrak{M}(\mathbf{v},\mathbf{w})$ consists of the framed torsion free sheaves (E,φ) such that the isomorphism class of $H^1(\mathbf{P}^2, E(-1))$, as a representation of Γ, is $\mathbf{v}$. Each $\mathfrak{M}(\mathbf{v},\mathbf{w})$, if it is nonempty, inherits a holomorphic symplectic and hyper-Kähler structures from $\mathfrak{M}(n,r)$.

Arbitrary quiver varieties of affine types with complex parameter equal to 0 are some $\mathfrak{M}(\mathbf{v},\mathbf{w})$. The identification with original definition was implicitly given in [29]. It was independently rediscovered by Lusztig [24]. See also [42]. Arbitrary quiver varieties of affine types with *nonzero* complex parameter are also important in representation theory [12], though we do not discuss here. Original definition of the varieties was given in terms of quivers. Later these were identified with framed moduli spaces of instantons on a noncommutative deformation $\mathbf{R}^4$ [36] or those of torsion free sheaves on a noncommutative deformation of $\mathbf{P}^2$ [18, 1].

4. Stratification and fibers of π

This technical section will be used to state character formulas later. A reader who only want to know only a *rough* picture can be skip this section.

We have the following stratification of $(S^n(\mathbf{C}^2))^\Gamma$ and its higher rank analog $\mathfrak{M}_0(n,r)^\Gamma$. The space $\mathfrak{M}_0(n,r)^\Gamma$ also decompose as

$$\begin{aligned} (S^n(\mathbf{C}^2))^\Gamma &= \bigsqcup_{m\le n} S^m_\lambda(\mathbf{C}^2/\Gamma), \\ \mathfrak{M}_0(n,r)^\Gamma &= \bigsqcup_{\substack{\mathbf{v}^0,\lambda \\ m+|\mathbf{v}^0|\le n}} \mathfrak{M}_0^{\mathrm{reg}}(\mathbf{v}^0,\mathbf{w}) \times S^m_\lambda(\mathbf{C}^2/\Gamma), \end{aligned} \tag{4.1}$$

where $\mathfrak{M}_0^{\mathrm{reg}}(\mathbf{v}^0,\mathbf{w})$ is defined exactly as above (it is possibly an empty set), $|\mathbf{v}^0|$ is the dimension of $\mathbf{v}^0$ as a complex vector space, $\lambda = (\lambda_1,\dots,\lambda_r)$ is a partition of m and

$$S^m_\lambda(\mathbf{C}^2/\Gamma) = \left\{ \sum_{i=1}^r \lambda_i[x_i] \in S^m(\mathbf{C}^2/\Gamma) \,\middle|\, x_i \ne 0 \text{ and } x_i \ne x_j \text{ for } i \ne j \right\}.$$

The differences $n-m$ and $n-(m+|\mathbf{v}^0|)$ are the multiplicity of the cycle at the origin 0.

Now it becomes clear that the case $(S^n(\mathbf{C}^2))^\Gamma$ is the special case of $\mathfrak{M}_0(n,r)^\Gamma$ with $\mathbf{w} = \rho_0$, $\mathbf{v}^0 = 0$. So from now, we only consider the second case.

For $x \in \mathfrak{M}_0(n,r)^\Gamma$, let $\mathfrak{M}(\mathbf{v},\mathbf{w})_x$ be the inverse image $\pi^{-1}(x)$ in $\mathfrak{M}(\mathbf{v},\mathbf{w})$. The most important one is the central fiber, i.e., the fiber over

$$x = (W \otimes_{\mathbf{C}} \mathcal{O}_{\mathbf{P}^2}, \varphi, 0).$$

In this case, we use the special notation $\mathfrak{L}(\mathbf{v},\mathbf{w})$. It is known that this is a Lagrangian subvariety of $\mathfrak{M}(\mathbf{v},\mathbf{w})$. Suppose that $x=(E_0,\varphi,C)$ is contained in the stratum $\mathfrak{M}_0^{\mathrm{reg}}(\mathbf{v}^0,\mathbf{w})\times S^m_\lambda(\mathbf{C}^2/\Gamma)$. Then the fiber $\mathfrak{M}(\mathbf{v},\mathbf{w})_x$ is a pure dimensional subvariety in $\mathfrak{M}(\mathbf{v},\mathbf{w})$, which is a product of $\mathfrak{L}(\mathbf{v}_s,\mathbf{w}_s)$ and copies of punctual Hilbert schemes $\mathrm{Hilb}_0^{\lambda_i}(\mathbf{C}^2)$ for some $\mathbf{v}_s$, $\mathbf{w}_s$. The proof of this statement in [27, §6], [31, §3] was given only when $m=0$ and explained in terms of quivers, so we give more direct argument in our situation. The fiber $\mathfrak{M}(\mathbf{v},\mathbf{w})_x$ parametrises Γ-invariant subsheaves E of E_0 such that $\left[H^1(\mathbf{P}^2,E(-1))\right]=\mathbf{v}$ and $\operatorname{Supp}E_0/E=C$. Equivalently, it parametrises Γ-equivariant 0-dimensional quotients $E_0\to Q$ such that $\left[H^0(\mathbf{P}^2,Q)\right]=\mathbf{v}-\mathbf{v}^0$ and $\operatorname{Supp}Q=C$. Such quotients depend only on a local structure on E_0, so we can replace E_0 by $W_s\otimes_{\mathbf{C}}\mathcal{O}_{\mathbf{P}^2}$, where W_s is the fiber of E_0 at the origin considered as a representation of Γ. The isomorphism class $\mathbf{w}_s$ of W_s is given by $\mathbf{w}_s=\mathbf{w}-\mathbf{C}\mathbf{v}^0$, where $\mathbf{C}$ is the class of the virtural representation $\bigwedge^0 Q-\bigwedge^1 Q+\bigwedge^2 Q=2\rho_0-Q$, and $\mathbf{C}\mathbf{v}^0$ means the tensor product $\mathbf{C}\otimes\mathbf{v}^0$. Therefore it becomes clear now that we have

$$\mathfrak{M}(\mathbf{v},\mathbf{w})_x\cong\mathfrak{L}(\mathbf{v}-\mathbf{v}^0-m\delta,\mathbf{w}_s)\times\prod_i\mathrm{Hilb}_0^{\lambda_i}(\mathbf{C}^2),$$

where δ is considered as the class of the regular representation, and $\mathrm{Hilb}_0^{\lambda_i}(\mathbf{C}^2)$ is the punctural Hilbert scheme, i.e., the inverse image of $\lambda_i[0]$ by the Hilbert-Chow morphism $\pi\colon \mathrm{Hilb}^{\lambda_i}(\mathbf{C}^2)\to S^{\lambda_i}(\mathbf{C}^2)$. The punctural Hilbert schemes are known to be irreducible, thus $\mathfrak{M}(\mathbf{v},\mathbf{w})_x$ is pure-dimensional if and only if $\mathfrak{L}(\mathbf{v}-\mathbf{v}^0-m\delta,\mathbf{w}_s)$ is so. But the latter statement is known [27, §5].

5. A geometric construction of the affine Lie algebra

After writing [21], the author tried to use this generalized ADHM description to study these varieties $\mathfrak{M}(\mathbf{v},\mathbf{w})$. But it turned out to be not so easy as he had originally hoped. When he struggled the problem, he heard a talk by Lusztig in ICM 90 Kyoto on a construction of canonical bases by using quivers. Lusztig's construction [23] was motivated by Ringel's construction [37] of the upper half part of the quantized enveloping algebra via the Hall algebra. The author thought that this construction should be useful to attack the problem. Two years later, he began to understand the picture. Quiver varieties $\mathfrak{M}(\mathbf{v},\mathbf{w})$ are, very *roughly*, cotangent bundles of varieties used by Ringel and Lusztig, and similar construction is possible [27]. A little later, he graduately realized that quiver varieties are also similar to cotangent bundles of flag varieties and the map π is an analog of Springer resolution. These varieties had been used to give geometric constructions of Weyl groups (Springer representations) and affine Hecke algebras (Deligne-Langlands conjecture). (See a beautifully written text book by N. Chriss and V. Ginzburg [8] and the references therein for these matereial.) The technique is the convolution product (see below) and works quite general. So he (and some others) conjectured that these construction should be adapted to quiver varieties. This conjecture turned

out to be true [28, 31]. We explain the constructions in this and next sections. The relation between our constructions and Ringel-Lusztig construction was explained in [30] and will not be reproduced here.

5.1. Convolution algebra

We apply the theory of the convolution algebra to varieties introduced in the previous sections to obtain the universal enveloping algebra $\mathbf{U}(\widehat{\mathfrak{g}})$ of the affine algebra $\widehat{\mathfrak{g}}$.

We continue to fix a representation W of Γ and denote by $\mathbf{w}$ its isomorphism class. For the Γ-fixed point set of Hilbert schemes, studied in §2, W is the trivial representation.

We introduce the following notation:

$$\mathfrak{M}(\mathbf{w}) \stackrel{\text{def.}}{=} \bigsqcup_n \mathfrak{M}(n,r)^\Gamma = \bigsqcup_{\mathbf{v}} \mathfrak{M}(\mathbf{v},\mathbf{w}), \qquad \mathfrak{L}(\mathbf{w}) \stackrel{\text{def.}}{=} \bigsqcup_{\mathbf{v}} \mathfrak{L}(\mathbf{v},\mathbf{w}),$$
$$\mathfrak{M}_0(\infty,\mathbf{w}) \stackrel{\text{def.}}{=} \bigcup_n \mathfrak{M}_0(n,r)^\Gamma.$$

The first and second are disjoint union. For the last, we use the inclusion $\mathfrak{M}_0(n,r)^\Gamma \subset \mathfrak{M}_0(n',r)^\Gamma$ for $n \le n'$ given by

$$(E,\varphi,C) \mapsto (E,\varphi,C+(n'-n)0)\,.$$

For $r=1$, these are

$$\bigsqcup_n \left(\mathrm{Hilb}^n(\mathbf{C}^2)\right)^\Gamma, \quad \bigcup_n (S^n\mathbf{C}^2)^\Gamma = \bigcup_n S^n(\mathbf{C}^2/\Gamma),$$

where the inclusion $S^n(\mathbf{C}^2/\Gamma) \subset S^{n'}(\mathbf{C}^2/\Gamma)$ is given by adding $(n'-n)0$ as above.

Rigorously speaking, we cannot study $\mathfrak{M}(\mathbf{w})$ and $\mathfrak{M}_0(\infty,\mathbf{w})$ directly since they are infinite dimensional. We need to work individual spaces $\mathfrak{M}(\mathbf{v},\mathbf{w})$, $\mathfrak{M}_0(n,r)^\Gamma$. But we use those spaces as if they are finite dimensional spaces for a notational convenience.

We consider the fiber product

$$Z(\mathbf{w}) \stackrel{\text{def.}}{=} \mathfrak{M}(\mathbf{w}) \times_{\mathfrak{M}_0(\infty,\mathbf{w})} \mathfrak{M}(\mathbf{w}).$$

It consists of pairs (E,φ), (E',φ') such that

1. $E^{\vee\vee} \cong E'^{\vee\vee}$
2. $\operatorname{Supp} E^{\vee\vee}$ and $\operatorname{Supp} E'^{\vee\vee}$ are equal in the complement of the origin.

The multiplicities of $\operatorname{Supp} E^{\vee\vee}$ and $\operatorname{Supp} E'^{\vee\vee}$ at the origin may be different since we consider the inclusion above.

One can show that this is a lagrangian subvariety in $\mathfrak{M}(\mathbf{w}) \times \mathfrak{M}(\mathbf{w})$. (The same remark as $\mathfrak{M}(\mathbf{v},\mathbf{w})_x$ in §4 applies here also.) Let us consider its top degree Borel-Moore homology group

$$H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C}).$$

More precisely, it is the subspace of

$$\prod_{\mathbf{v}^1,\mathbf{v}^2} H_{\mathrm{top}}(Z(\mathbf{w}) \cap (\mathfrak{M}(\mathbf{v}^1,\mathbf{w}) \times \mathfrak{M}(\mathbf{v}^2,\mathbf{w})), \mathbf{C})$$

consisting of elements $(F_{\mathbf{v}^1,\mathbf{v}^2})$ such that

1. for fixed $\mathbf{v}^1$, $F_{\mathbf{v}^1,\mathbf{v}^2} = 0$ for all but finitely many choices of $\mathbf{v}^2$,
2. for fixed $\mathbf{v}^2$, $F_{\mathbf{v}^1,\mathbf{v}^2} = 0$ for all but finitely many choices of $\mathbf{v}^1$.

The degree top depends on $\mathbf{v}^1$, $\mathbf{v}^2$, but we supress the dependency for brevity.

Let us consider the convolution product

$$*\colon H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C}) \otimes H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C}) \to H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C})$$

given by

$$c * c' = p_{13*}\left(p_{12}^*(c) \cap p_{23}^*(c')\right),$$

where p_{ij} is the projection from the triple product $\mathfrak{M}(\mathbf{w}) \times \mathfrak{M}(\mathbf{w}) \times \mathfrak{M}(\mathbf{w})$ to the double product $\mathfrak{M}(\mathbf{w}) \times \mathfrak{M}(\mathbf{w})$ of the ith and jth factors. More detail for the definition of the convolution product, say p_{12}^*, $\cap$, is explained in [8], but we want to emphasize one point. The statement that the result $c * c'$ has top degree is the consequence of $\dim Z(\mathbf{w}) = \frac{1}{2}\dim \mathfrak{M}(\mathbf{w}) \times \mathfrak{M}(\mathbf{w})$. Although we are considering $Z(\mathbf{w})$ having infinitely many connected components, the convolution is well-defined and $H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C})$ is an associative algebra with unit, thanks to the above definition of the subspace in the direct product.

For $x \in \mathfrak{M}_0(\infty,\mathbf{w})$, let $\mathfrak{M}(\mathbf{w})_x$ be the inverse image $\pi^{-1}(x)$ in $\mathfrak{M}(\mathbf{w})$. We consider the top degree homology group

$$H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_x,\mathbf{C})$$

which is the usual direct sum of $H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_x \cap \mathfrak{M}(\mathbf{v},\mathbf{w}),\mathbf{C})$ (unlike the case of $Z(\mathbf{w})$). The convolution product makes this space into a module of $H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C})$.

Theorem 5.1. *Let $\mathbf{U}(\widehat{\mathfrak{g}})$ be the universal enveloping algebra of the untwisted affine algebra $\widehat{\mathfrak{g}}$ corresponding to Γ. (**NB**: not a 'quantum' version). There exists an algebra homomorphism*

$$\mathbf{U}(\widehat{\mathfrak{g}}) \to H_{\mathrm{top}}(Z(\mathbf{w}),\mathbf{C}).$$

Furthermore, if we consider $H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_x,\mathbf{C})$ as a $\mathbf{U}(\widehat{\mathfrak{g}})$-module via the homomorphism, it is an irreducible integrable highest weight representation and the direct summands $H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_x \cap \mathfrak{M}(\mathbf{v},\mathbf{w}),\mathbf{C})$ are weight spaces.

This theorem was essentially proved in [27] with a modification for general x mentioned above.

The highest weight of $H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_x,\mathbf{C})$ and weights of $H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_x \cap \mathfrak{M}(\mathbf{v},\mathbf{w}),\mathbf{C})$ are determined explicitly in terms of $\mathbf{v}$, $\mathbf{w}$ and the stratum to which x belongs. For example, in the case of the central fiber $\mathfrak{L}(\mathbf{v},\mathbf{w})$, the highest weight is $\mathbf{w}$, considered as a dominant integral weight as $\mathbf{w} = \sum_i w_i\Lambda_i$, where w_i is the ρ_i component of

$\mathbf{w}$, and Λ_i is the ith fundamental weight. Here we use the identification of the irreducible representation ρ_i and a vertex of the affine Dynkin diagram given by McKay correspondence. The weight of $H_{\text{top}}(\mathfrak{L}(\mathbf{v},\mathbf{w}),\mathbf{C})$ is $\mathbf{w}-\mathbf{v}$, where $\mathbf{v}=\sum_i v_i\alpha_i$ with the ρ_i-component v_i of $\mathbf{v}$ and ith simple root α_i. The highest weight vector is the fundamental class $[\mathfrak{L}(0,\mathbf{w})]$, where $\mathfrak{M}(0,\mathbf{w})=\mathfrak{L}(0,\mathbf{w})$ consists of a single point $E=W\otimes_{\mathbf{C}}\mathcal{O}_{\mathbf{P}^2}$.

For the case studied in §2, $\mathbf{w}$ is the 0th fundamental weight Λ_0. The corresponding integrable highest weight representation is called the *basic* representation in literature. If we vary $\mathbf{w}$, we get all integrable highest weight representations as $\bigoplus_{\mathbf{v}} H_{\text{top}}(\mathfrak{L}(\mathbf{v},\mathbf{w}),\mathbf{C})$. It is worth while remarking that this is an extension of (1.1) since the Cartan subalgebra $\mathfrak{h}$ is naturally contained in the Cartan subalgebra. Furthermore, the finite dimensional Lie algebra $\mathfrak{g}$ is embedded in the basic representation, and we get

$$\mathfrak{g}\cong\bigoplus_{v_0=1} H_{\text{top}}(M(\mathbf{v}),\mathbf{C}),$$

where v_0 is the ρ_0-isotropic component of $\mathbf{v}$. This is an extension of (1.1), mentioned before. In fact, it is easy to see that if $\mathbf{v}$ is not δ, then $M(\mathbf{v})$ is either empty, or a single point. The latter holds if and only if $\mathbf{v}$, considered as an element of $\mathfrak{h}^*$ by removing v_0, is a root of $\mathfrak{g}$.

If we fix $\mathbf{w}$ and vary x, we still obtain various integrable highest weight representations. The highest weight of $H_{\text{top}}(\mathfrak{M}(\mathbf{w})_x,\mathbf{C})$ is $\mathbf{w}-\mathbf{v}^0-m\delta$, where $\mathbf{v}^0$, m are determined by x as in §4, and $\mathbf{w}$, $\mathbf{v}^0$ are considered as weights as above. The weight of $H_{\text{top}}(\mathfrak{M}(\mathbf{w})_x\cap\mathfrak{M}(\mathbf{v},\mathbf{w}),\mathbf{C})$ is equal to $\mathbf{w}-\mathbf{v}$. All of their highest weights are less than or equal to $\mathbf{w}$ with respect to the dominance order. In particular, when $\mathbf{w}=\Lambda_0$, those have highest weights $\Lambda_0-n\delta$ for some $n\in\mathbf{Z}_{\geq 0}$. They are essentially isomorphic to the basic representation.

We explain how the algebra homomorphism $\mathbf{U}(\widehat{\mathfrak{g}})\to H_{\text{top}}(Z(\mathbf{w}),\mathbf{C})$ is defined. It is enough to define the image of Chevalley generators e_i, f_i, h_i $(i\in I)$, d of $\mathbf{U}(\widehat{\mathfrak{g}})$ (and check the defining relations). The images of h_i and d are multiples of fundamental classes of diagonales in $\mathfrak{M}(\mathbf{w})\times\mathfrak{M}(\mathbf{w})$. More precisely, the multiple is determined so that the weight of $H_{\text{top}}(\mathfrak{M}(\mathbf{w})_x\cap\mathfrak{M}(\mathbf{v},\mathbf{w}),\mathbf{C})$ is equal to $\mathbf{w}-\mathbf{v}$. The image of e_i is the fundamental classe of the so-called 'Hecke correspondence':

$$\bigsqcup_{\mathbf{v}}\{((E,\varphi),(E',\varphi'))\in\mathfrak{M}(\mathbf{v},\mathbf{w})\times\mathfrak{M}(\mathbf{v}+\rho_i,\mathbf{w})\mid E\subset E'\}. \tag{5.1}$$

It is known that each component is a nonsingular lagrangian subvariety of $\mathfrak{M}(\mathbf{v},\mathbf{w})\times\mathfrak{M}(\mathbf{v}+\rho_i,\mathbf{w})$. Hence it is an irreducible component of $Z(\mathbf{w})$. The image of f_i is given by swapping the first and second factors, up to sign.

As an application of the above construction, we get a base of $H_{\text{top}}(\mathfrak{M}(\mathbf{w})_x,\mathbf{C})$ indexed by the irreducible components of $\mathfrak{M}(\mathbf{w})_x$. It has a structure of the crystal in the sense of Kashiwara, and is isomorphic to the crystal of the corresponding integrable highest weight module of the quantum affine algebra by Kashiwara-Saito [19, 38]. (See also [32] for a different proof.) However, the base itself is different from the specialization of the canonical (= global crystal) base of the quantum

affine algebra module at $q = 1$. A counter example was found in [19]. The base given by irreducible components is named *semicanonical base* by Lusztig [25].

5.2. Lower degree homology groups

The construction in the previous subsection, in fact, gives us a structure of a representation of $\widehat{\mathfrak{g}}$ on

$$H_{\mathrm{top}-d}(\mathfrak{M}(\mathbf{w})_x, \mathbf{C})$$

for each fixed integer d. It is an integrable representation, and decomposes into irreducible representations. The multiplicity formula can be expressed in terms of the intersection cohomology thanks to Beilinson-Bernstein-Deligne-Gabber's decomposition theorem [3] applied to the morphism $\pi\colon \mathfrak{M}(\mathbf{w}) \to \mathfrak{M}_0(\infty, \mathbf{w})$. In our situation, π is a *semi-small* morphism, i.e., the restriction of π to the inverse image of the stratum (4.1) is a topological fiber bundle, and

$$2 \dim \mathfrak{M}(\mathbf{w})_x \le \operatorname{codim} \mathcal{O}_x,$$

where $\mathcal{O}_x$ is the stratum containing x. Then as observed by Borho-MacPherson [4], the decomposition theorem is simplified. We introduce several notation to describe the formula. We choose a point y from each stratum in (4.1). We denote the stratum containing y by $\mathcal{O}_y$. Let $IC(\mathcal{O}_y)$ is the intersection homology complex of $\mathcal{O}_y$ with respect to the trivial local system.

Theorem 5.2. *We have the following decomposition as a representation of* $\widehat{\mathfrak{g}}$:

$$H_{\mathrm{top}-d}(\mathfrak{M}(\mathbf{w})_x, \mathbf{C}) = \bigoplus_y H^{d+\dim \mathcal{O}_x}(i_x^! IC(\mathcal{O}_y)) \otimes H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_y, \mathbf{C}),$$

where $i_x\colon \{x\} \to \mathfrak{M}_0(\infty, \mathbf{w})$ *is the inclusion. Here* $\widehat{\mathfrak{g}}$ *acts trivially on the first factor of the right hand side.*

For a general semi-small morphism, we may have an intersection homology complex with respect to a *nontrivial* local system in the decomposition. In order to show that such a summand does not appear, the fact that $H_{\mathrm{top}}(\mathfrak{M}(\mathbf{w})_y, \mathbf{C})$ is a highest weight module plays a crucial role (see [31] for detail).

Note also that when $\mathbf{w} = \rho_0$, the closure of each stratum is a symmetric product of $\mathbf{C}^2/\Gamma$ (4.1). In particular, they only have finite quotient singularities, and their intersection homology complex are equal to the constant sheaf. Therefore our formula are simplified. One finds that $H_*(\mathfrak{L}(\mathbf{w}), \mathbf{C})$ is isomorphic to the so-called 'Fock space'. Later we will show that the total homology $H_*(\mathfrak{L}(\mathbf{w}), \mathbf{C})$ has a structure of a representation of a specialized quantum toroidal algebra in the next section. Then this observation generalizes a result [41, 39] for type $A_n^{(1)}$ to untwisted affine Lie algebras of type ADE.

6. Equivariant K-theory and quantum toroidal algebras

In this section, we replace the top degree homology group H_{top} in the previous section by equivariant K-groups. Then we obtain a geometric construction of a quantum toroidal algebra $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$. It is a q–analog of the loop algebra $\mathbf{L}\widehat{\mathfrak{g}} = \widehat{\mathfrak{g}} \otimes_{\mathbf{C}} \mathbf{C}[z, z^{-1}]$ of the affine algebra $\widehat{\mathfrak{g}}$. Since $\widehat{\mathfrak{g}}$ is already a (central extension of) loop alebra of $\mathfrak{g}$, $\mathbf{L}\widehat{\mathfrak{g}}$ is a 'double-loop' algebra of $\mathfrak{g}$. The quantum toroidal algebra is defined by replacing $\mathfrak{g}$ by $\widehat{\mathfrak{g}}$ in the so-called Drinfeld realization of the quantum loop algebra $\mathbf{U}_q(\mathbf{L}\mathfrak{g})$, which is a subquotient of the quantum affine algebra $\mathbf{U}_q(\widehat{\mathfrak{g}})$, defined by Drinfeld, Jimbo.

Let $G_{\mathbf{w}} = \operatorname{Aut}_\Gamma(W)$ be the group of automorphisms of the Γ-module W. If w_i is the multiplicity of ρ_i in W, we have $G_{\mathbf{w}} \cong \prod_i \mathrm{GL}_{w_i}(\mathbf{C})$. We have a natural action of $G_{\mathbf{w}}$ on $\mathfrak{M}(\mathbf{w})$ and $\mathfrak{M}_0(\infty, \mathbf{w})$ by the change of the framing:

$$\varphi \mapsto g \circ \varphi, \qquad g \in G_{\mathbf{w}}.$$

The projective morphism $\mathfrak{M}(\mathbf{w}) \to \mathfrak{M}_0(\infty, \mathbf{w})$ is equivariant.

Let $\mathbf{C}^*$ act on $\mathbf{C}^2$ by $t \cdot (x, y) = (tx, ty)$. It extends to an action on $\mathbf{P}^2$, where it acts trivially on ℓ_∞. Note that this action commutes with the Γ-action. Then we have a natural induced $\mathbf{C}^*$-action on $\mathfrak{M}(\mathbf{w})$ and $\mathfrak{M}_0(\infty, \mathbf{w})$ so that the projection π is equivariant. Combining two actions we have an action of $G_{\mathbf{w}} \times \mathbf{C}^*$ on $\mathfrak{M}(\mathbf{w})$ and $\mathfrak{M}_0(\infty, \mathbf{w})$. (This action is different from the action studied in [31] for type $A_1^{(1)}$. We need to change the definition of $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$ in that paper to apply the result in this section. This comes from the umbiguity of the definition of a q-analog of the Cartan matrix.)

Let $K^{G_{\mathbf{w}} \times \mathbf{C}^*}(Z(\mathbf{w}))$ be the equivariant K-homology group of $Z(\mathbf{w})$ with respect to the above $G_{\mathbf{w}} \times \mathbf{C}^*$-action. (More precisely, it should be defined as a subspace of the direct product as in the case of homology groups.) It is a module over the representation ring $R(G_{\mathbf{w}} \times \mathbf{C}^*) = \mathbf{Z}[q, q^{-1}] \otimes_{\mathbf{Z}} R(G_{\mathbf{w}})$, where q is the natural 1-dimensional representation of $\mathbf{C}^*$. The convolution product makes $K^{G_{\mathbf{w}} \times \mathbf{C}^*}(Z(\mathbf{w}))$ into a $R(G_{\mathbf{w}} \times \mathbf{C}^*)$-algebra. We divide its torsion part over $\mathbf{Z}[q, q^{-1}]$ and denote it by $K^{G_{\mathbf{w}} \times \mathbf{C}^*}(Z(\mathbf{w}))/\operatorname{torsion}$. (It is conjectured that the torsion is, in fact, 0.)

Theorem 6.1. *There exists a $\mathbf{Z}[q, q^{-1}]$-algebra homomorphism*

$$\mathbf{U}_q^{\mathbf{Z}}(\mathbf{L}\widehat{\mathfrak{g}}) \to K^{G_{\mathbf{w}} \times \mathbf{C}^*}(Z(\mathbf{w}))/\operatorname{torsion},$$

where $\mathbf{U}_q^{\mathbf{Z}}(\mathbf{L}\widehat{\mathfrak{g}})$ is a certain $\mathbf{Z}[q, q^{-1}]$-subalgebra (conjecturally an integral form) of $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$.

The definition of the homomorphism is similar to the case of homology groups. The image of the q-analog of $e_i \otimes z^r$ is given by a natural line bundles on the Hecke correspondence (5.1) whose fiber at $((E, \varphi), (E', \varphi'))$ is $H^0(E'/E)^{\otimes r}$.

Let us explain how we can use this algebra homomorphism to study representations of specialized quantum toroidal algebra $\mathbf{U}_\varepsilon(\mathbf{L}\widehat{\mathfrak{g}}) = \mathbf{U}_q^{\mathbf{Z}}(\mathbf{L}\widehat{\mathfrak{g}})|_{q=\varepsilon}$, where ε is a nonzero complex number which may or may not be a root of unity. A natural

generalization of finite dimensional representations of $\mathbf{U}_\varepsilon(\mathbf{L}\mathfrak{g})$ are l–integrable representations. (See [31] for the definition.) The Drinfeld-Chari-Pressley classification [11, 7] of irreducible finite dimensional reprentation of $\mathbf{U}_\varepsilon(\mathbf{L}\mathfrak{g})$ has a natural analog in $\mathbf{U}_\varepsilon(\mathbf{L}\widehat{\mathfrak{g}})$: Irreducible l–integrable representations of $\mathbf{U}_\varepsilon(\mathbf{L}\widehat{\mathfrak{g}})$ are parametrized by I–tuple of polynomials $P_i(u)$ with $P_i(0) = 1$, where I is the set of vertices of the affine Dynkin diagram.

Irreducible representations are obtained in the following way. Let us consider the equivariant homology $K^{G_\mathbf{w}\times\mathbf{C}^*}(\mathfrak{L}(\mathbf{w}))$, which is without torsion. It is a module of $\mathbf{U}_q^\mathbf{Z}(\mathbf{L}\widehat{\mathfrak{g}})$ and called a *universal standard module.* For a semisimple element $(s,\varepsilon) \in G_\mathbf{w} \times \mathbf{C}^*$, we consider the evaluation homomorphism $R(G_\mathbf{w} \times \mathbf{C}^*) \to \mathbf{C}$. Then the specialization

$$K^{G_\mathbf{w}\times\mathbf{C}^*}(\mathfrak{L}(\mathbf{w})) \otimes_{R(G_\mathbf{w}\times\mathbf{C}^*)} \mathbf{C}$$

is a representation of $\mathbf{U}_\varepsilon(\mathbf{L}\widehat{\mathfrak{g}})$. This is called a *standard module.* It has a unique irreducible quotient, and the associated polynomials are the characteristic polynomials of components of s. (Recall $G_\mathbf{w} = \prod_{i\in I} \mathrm{GL}_{w_i}(\mathbf{C})$.)

In order to state character formulas, which is very similar to Theorem 5.2, we need a little more notation. Let A be the Zariski closure of powers of (s,ε) in $G_\mathbf{w} \times \mathbf{C}^*$. Let $\mathfrak{M}(\mathbf{w})^A$, $\mathfrak{M}_0(\infty,\mathbf{w})^A$, $Z(\mathbf{w})^A$ be the fixed point sets. We have a chain of natural algebra homomorphisms

$$\begin{aligned} K^{G_\mathbf{w}\times\mathbf{C}^*}(Z(\mathbf{w}))_{R(G_\mathbf{w}\times\mathbf{C}^*)}\mathbf{C} &\to K^A(Z(\mathbf{w}))_{R(A)}\mathbf{C} \\ &\to K(Z(\mathbf{w})^A) \otimes_\mathbf{Z} \mathbf{C} \to H_*(Z(\mathbf{w})^A, \mathbf{C}), \end{aligned}$$

where the first one is induced by the inclusion $A \subset G_\mathbf{w} \times \mathbf{C}^*$, the second one is given by the localization theorem in the equivariant K-theory, and the last one is the Chern character homomorphism. (In fact, we need 'twists' for the last two. See [8] for detail.)

There exists a natural stratification of $\mathfrak{M}_0(\infty,\mathbf{w})^A$ similar to (4.1). We choose a point y in each stratum and denote by $\mathcal{O}_y$ the stratum containing y. If $\mathfrak{M}(\mathbf{w})^A_y$ denotes the inverse image of y in $\mathfrak{M}(\mathbf{w})^A$ under π, the homology group $H_*(\mathfrak{M}(\mathbf{w})^A_y, \mathbf{C})$ is a representation of $H_*(Z(\mathbf{w})^A, \mathbf{C})$, and hence that of $\mathbf{U}_\varepsilon(\mathbf{L}\widehat{\mathfrak{g}})$. (When $y = 0$, it is the standard module.) Analog of Theorem 5.2 is the following:

Theorem 6.2. *We have the following in the Grothendieck group of the abelian category of l–integrable representations of $\mathbf{U}_\varepsilon(\mathbf{L}\widehat{\mathfrak{g}})$*

$$H_*(\mathfrak{M}(\mathbf{w})^A_x, \mathbf{C}) = \sum_y H^*(i_x^! IC(\mathcal{O}_y)) \otimes L_y,$$

where L_y is the unique irreducible quotient of $H_*(\mathfrak{M}(\mathbf{w})^A_y, \mathbf{C})$.

(The right hand side is an infinite sum, so we must understand it in an appropriate way. But it should be clear how it can be done.)

In this case, we apply the decomposition theorem to $\mathfrak{M}(\mathbf{w})^A \to \mathfrak{M}_0(\infty,\mathbf{w})^A$. It is *not* semi-small any more. So the degrees in the left and righ hand sides do not have clear relations.

Remarks 6.3. (1) We stated our results for the affine Lie algebra $\widehat{\mathfrak{g}}$ and the quantum toroidal algebra $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$. But they also hold for the finite dimensional Lie algebra $\widehat{\mathfrak{g}}$ and the quantum loop algebra $\mathbf{U}_q(\mathbf{L}\mathfrak{g})$, if we impose the condition $v_0 = w_0 = 0$. It is known that $\mathbf{U}_q(\mathbf{L}\mathfrak{g})$ is a Hopf algebra (since it is a subquotient of the quantum affine algebra), and the standard modules are isomorphic to tensor products of l–fundamental representations when ε is not a root of unity [43]. Here l–fundamental representations are irreducible representations corresponding to $\mathbf{w} = \rho_i$. In particular, the tensor product decomposition in the representation ring can be expressed in terms of intersection homology groups.

(2) Our remaining tasks are computing dimensions of $H^*(i_x^! IC(\mathcal{O}_y))$ appearing in Theorems 5.2, 6.2. In [33, 34] we gave a purely combinatorial algorithm to compute them. This algorithm can be made into a computer program. The algorithm was stated for the quantum loop algebra $\mathbf{U}_q(\mathbf{L}\mathfrak{g})$, but works also for $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$. This means that for any given stratum $\mathcal{O}_y$, $\dim H^*(i_x^! IC(\mathcal{O}_y))$ is, in principle, computable. However, it is practically difficult to compute because we need lots of memory. And, for $\mathbf{U}_q(\mathbf{L}\widehat{\mathfrak{g}})$, the summation is infinite. So having an algorithm to compute each term is not a strong statement. It is desirable to have an alternative method to compute them. That has be done in some special classes of representations [35].

References

[1] V. Baranovsky, V. Ginzburg and A. Kuznetsov, *Quiver varieties and a noncommutative* $\mathbb{P}^2$, preprint, arXiv:math.AG/0103068.

[2] W. Barth, C. Peters and A. Van de Ven, *Compact complex surfaces*, A Series of Modern Surveys in Math. **4**, Springer-Verlag, 1984.

[3] A. Beilinson, J. Bernstein and P. Deligne, *Faisceaux pervers*, Astérisque **100** (1982).

[4] W. Borho and R. MacPherson, *Partial resolutions of nilpotent varieties*, Astérisque **101–102** (1983), 23–74.

[5] T. Bridgeland, A. King and M. Reid, *The McKay correspondence as an equivalence of derived categories*, J. Amer. Math. Soc. **14** (2001), 535–554.

[6] H. Cassens and P. Slodowy, *Kleinian singularities and quivers*, in *Singularities (Oberwolfach, 1996)*, 263–288, Progr. Math., **162**, Birkhäuser, Basel, 1998.

[7] V. Chari and A. Pressley, *Quantum affine algebras and their representations*, in *Representations of groups (Banff, AB, 1994)*, Amer. Math. Soc., Providence, RI, 1995, pp. 59–78.

[8] N. Chriss and V. Ginzburg, *Representation theory and complex geometry*, Progress in Math., Birkhäuser, 1997.

[9] W. Crawley-Boevey, *Geometry of the moment map for representations of quivers*, Compositio Math. **126** (2001), 257–293.

[10] S.K. Donaldson and P.B. Kronheimer, *The geometry of four-manifold*, Oxford Math. Monographs, Oxford Univ. Press, 1990.

[11] V.G. Drinfel'd, *A new realization of Yangians and quantized affine algebras*, Soviet math. Dokl. **32** (1988), 212–216.

[12] P. Etingof and V. Ginzburg, *Symplectic reflection algebras, Calogero-Moser space, and deformed Harish-Chandra homomorphism*, Invent. Math. **147** (2002), 243–348.
[13] I.B. Frenkel, N. Jing and W. Wang, *Vertex representations via finite groups and the McKay correspondence*, Internat. Math. Res. Notices **2000**, no. 4, 195–222.
[14] G. Gonzalez-Sprinberg and J.L. Verdier, *Construction géometrique de la correspondence de McKay*, Ann. Sci. École Norm. Sup. **16** (1983), 409–449.
[15] M. Haiman, a talk at CRM, Université de Montréal, June 2002.
[16] D. Huybrechts and M. Lehn, *Stable pairs on curves and surfaces*, J. Algebraic Geom. **4** (1995), 67–104.
[17] Y. Ito and I. Nakamura, *Hilbert schemes and simple singularities*, in *Recent Trends in Algebraic Geometry* – EuroConference on Algebraic Geometry (Warwick, July 1996), ed. by K. Hulek and others, CUP, 1999, pp. 151–233.
[18] A. Kapustin, A. Kuznetsov and D. Orlov, *Noncommutative instantons and twistor transform*, Comm. Math. Phys. **221** (2001), 385–432.
[19] M. Kashiwara and Y. Saito, *Geometric construction of crystal bases*, Duke Math. **89** (1997), 9–36.
[20] P.B. Kronheimer, *The construction of ALE spaces as a hyper-Kähler quotients*, J. Differential Geom. **29** (1989) 665–683.
[21] P.B. Kronheimer and H. Nakajima, *Yang-Mills instantons on ALE gravitational instantons*, Math. Ann. **288** (1990), 263–307.
[22] A. Kuznetsov, *Quiver varieties and Hilbert schemes*, preprint, arXiv:math.AG/0101092.
[23] G. Lusztig, *Canonical bases arising from quantized enveloping algebras*, J. Amer. Math. Soc. **3** (1990), 447–498.
[24] ———, *On quiver varieties*, Adv. in Math. **136** (1998), 141–182.
[25] ———, *Semicanonical bases arising from enveloping algebras*, Adv. Math. **151** (2000), 129–139.
[26] J. McKay, *Graphs, singularities and finite groups*, Proc. Sympos. Pure Math. **37** Amer. Math. Soc. (1980), 183–186.
[27] H. Nakajima, *Instantons on ALE spaces, quiver varieties, and Kac-Moody algebras*, Duke Math. J. **76** (1994), 365–416.
[28] ———, *Quiver varieties and Kac-Moody algebras*, Duke Math. J. **91** (1998), no. 3, 515–560.
[29] ———, *Lectures on Hilbert schemes of points on surfaces*, Univ. Lect. Ser. **18**, AMS, 1999.
[30] ———, *Ebira tayoutai to Ryousi affine kan (Quiver varieties and quantum affine a lgebras)* (in Japanese), Suugaku **52** (2000), 337–359.
[31] ———, *Quiver varieties and finite dimensional representations of quantum affine algebras*, J. Amer. Math. Soc. **14** (2001), 145–238.
[32] ———, *Quiver varieties and tensor products*, Invent. Math., **146** (2001), 399–449.
[33] ———, *t-analogue of the q-characters of finite dimensional representations of quantum affine algebras*, in *Physics and Combinatorics*, Proceedings of the Nagoya 2000 International Workshop, World Scientific, 2001, 195–218.

[34] ———, *Quiver varieties and t–analogs of q–characters of quantum affine algebras*, preprint, arXiv:math.QA/0105173.
[35] ———, *t–analogs of q–characters of Kirillov-Reshetikhin modules of quantum affine algebras*, preprint, arXiv:math.QA/0204185.
[36] N. Nekrasov and A. Schwarz, *Instantons on noncommutative $\mathbf{R}^4$, and $(2,0)$ superconformal six-dimensional theory*, Comm. Math. Phys. **198** (1998), 689–703.
[37] C.M. Ringel, *Hall algebras and quantum groups*, Invent. Math. **101** (1990), 583–592.
[38] Y. Saito, *Geometric construction of crystal bases II*, preprint.
[39] Y. Saito, K. Takemura and D. Uglov, *Toroidal actions on level 1 modules of $U_q(\widehat{\mathrm{sl}}_n)$*, Transform. Groups **3** (1998), 75–102.
[40] P. Slodowy, *Simple singularities and simple algebraic groups*, Lecture Notes in Math. **815**, Springer, Berlin, 1980.
[41] M. Varagnolo and E. Vasserot, *Double-loop algebras and the Fock space*, Invent. Math. **133** (1998), 133–159.
[42] ———, *On the K-theory of the cyclic quiver variety*, Internat. Math. Res. Notices **1999**, no. 18, 1005–1028.
[43] ———, *Standard modules of quantum affine algebras*, Duke Math. J. **111** (2002), 509–533.
[44] W. Wang, *Hilbert schemes, wreath products, and the McKay correspondence*, preprint, arXiv:math.AG/9912104.

ICM 2002 · Vol. I · 439–448

Some Recent Transcendental Techniques in Algebraic and Complex Geometry*

Yum-Tong Siu†

Abstract

This article discusses the recent transcendental techniques used in the proofs of the following three conjectures. (1) The plurigenera of a compact projective algebraic manifold are invariant under holomorphic deformation. (2) There exists no smooth Leviflat hypersurface in the complex projective plane. (3) A generic hypersurface of sufficiently high degree in the complex projective space is hyperbolic in the sense that there is no nonconstant holomorphic map from the complex Euclidean line to it.

2000 Mathematics Subject Classification: 20C30, 20J05.
Keywords and Phrases: Plurigenera, Levi-Flat, Hyperbolicity.

1. Introduction

Since the use of function theory in the study of algebraic curves as Riemann surfaces about two hundred years ago, transcendental methods such as harmonic forms in Hodge theory and curvature in the theory of Chern-Weil have been very important tools in complex algebraic geometry. Since the nineteen sixties very powerful techniques in the estimates of $\bar{\partial}$, especially L^2 estimates and regularity techniques, have been extensively developed by C. B. Morrey, J. J. Kohn, L. Hörmander, *et al.* (To avoid too lengthy a bibliography here, we refer to [2],[4],[7],[10],[15],[16],[18] for references not listed here.) During the last two decades these new transcendental techniques have been increasingly used in complex algebraic geometry. The most noteworthy among them is J. J. Kohn's method of multiplier ideals for $\bar{\partial}$ estimates [7] which holds the promise of applicability to general partial differential equations and global geometry. Nadel [11] introduced multiplier ideal sheaves dual to Kohn's. A number of longstanding problems in algebraic and complex geometry hitherto beyond the reach of known methods have been solved by

*Research partially supported by a grant from the National Science Foundation.
†Department of Mathematics, Harvard University, Cambridge, MA 02138, USA. E-mail: siu@math.harvard.edu.

the new techniques of $\bar{\partial}$ estimates. On the other hand, demands of geometric applications motivate new approaches to $\bar{\partial}$ estimates. We will discuss here some recent results in the following three topics in algebraic and complex geometry obtained by the new transcendental methods. (1) Invariance of plurigenera. (2) Nonexistence of smooth Levi-flat hypersurface in $\mathbf{P}_2$. (3) Hyperbolicity of generic hypersurface of high degree in $\mathbf{P}_n$. Though topic (3) is only peripherally linked to $\bar{\partial}$ estimates, a long outstanding problem there is solved by some recent transcendental techniques.

2. Invariance of plurigenera

Let $\Delta_r = \{t \in \mathbf{C} \,|\, |t| < r\}$ and $\Delta = \Delta_1$. Denote by K_Y the canonical line bundle of a complex manifold Y. The m-genus of a compact complex manifold X is the complex dimension of $\Gamma(X, m\,K_X)$. By Hodge theory the 1-genus of a compact Kähler manifold is a topological invariant and therefore is invariant under holomorphic deformation. For the general m-genus there is the following conjecture on its invariance under holomorphic deformation for a compact Kähler manifold.

Conjecture 2.1 *(on Deformational Invariance of Plurigenera for Kähler Manifolds). Let $\pi : X \to \Delta$ be a holomorphic family of compact Kähler manifolds with fiber X_t. Then for any positive integer m the complex dimension of $\Gamma(X_t, m\,K_{X_t})$ is independent of t for $t \in \Delta$.*

Conjecture (2.1) has been verified in [20] when X is a family of compact projective algebraic manifolds.

Theorem 2.2 [20]. *Let $\pi : X \to \Delta$ be a holomorphic family of compact complex projective algebraic manifolds. Then for any integer $m \geq 1$ the complex dimension of $\Gamma(X_t, m\,K_{X_t})$ is independent of t for $t \in \Delta$.*

The main techniques to solve the problem were first introduced in [17] where for technical reasons the assumption of each fiber being of general type is added. Because of the upper semicontinuity of $\dim_{\mathbf{C}} \Gamma(X_t, m\,K_{X_t})$, the conjecture is equivalent to extending every element of $\Gamma(X_t, m\,K_{X_t})$ to $\Gamma(X, m\,K_X)$. We can assume $t = 0$. The idea of the main techniques stemmed from the following naive motivation. If one could write an element $s^{(m)}$ of $\Gamma(X_0, m\,K_{X_0})$ as a sum of terms, each of which is the product of an element $s^{(1)}$ of $\Gamma(X_0, K_{X_0})$ and an element $s^{(m-1)}$ of $\Gamma(X_0, (m-1)\,K_{X_0})$, then one can extend $s^{(m)}$ to an element of $\Gamma(X, m\,K_X)$ by induction on m. Of course, in general it is clearly impossible to so express $s^{(m)}$ as a sum of such products. However, one could successfully implement a modified form of this naive motivation, in which $s^{(1)}$ is only a local holomorphic section and $s^{(m-1)}$ is an element of $\Gamma(X_0, (m-1)\,K_{X_0} + E)$ instead of $\Gamma(X_0, (m-1)\,K_{X_0})$, where E is a sufficiently ample line bundle on X independent of m. The implementation of the modified form depends on the following two ingredients.

Proposition 2.3 *(Global Generation of Multiplier Ideal Sheaves). Let L be a holomorphic line bundle over an n-dimensional compact complex manifold Y with a Hermitian metric which is locally of the form $e^{-\xi}$ with ξ plurisubharmonic. Let $\mathcal{I}_\xi$ be the multiplier ideal sheaf of the Hermitian metric $e^{-\xi}$ (i.e., the sheaf consisting of all holomorphic function germs f with $|f|^2\,e^{-\xi}$ locally integrable). Let E be an ample holomorphic line bundle over Y such that for every point P of Y there are a*

finite number of elements of $\Gamma(Y,E)$ *which all vanish to order at least* $n+1$ *at* P *and which do not simultaneously vanish outside* P. *Then* $\Gamma(Y,\mathcal{I}_\xi\otimes(L+E+K_Y))$ *generates* $\mathcal{I}_\xi\otimes(L+E+K_Y)$ *at every point of* Y.

Proposition 2.4 *(Extension Theorem of Ohsawa-Takegoshi Type). Let* $\gamma: Y\to\Delta$ *be a projective algebraic family of compact complex manifolds. Let* $Y_0=\gamma^{-1}(0)$ *and let* n *be the complex dimension of* Y_0. *Let* L *be a holomorphic line bundle with a Hermitian metric* $e^{-\chi}$ *with* χ *plurisubharmonic. Then for* $0<r<1$ *there exists a positive constant* A_r *with the following property. For any holomorphic* L*-valued* n*-form* f *on* Y_0 *with* $\int_{Y_0}|f|^2e^{-\chi}<\infty$, *there exists a holomorphic* L*-valued* $(n+1)$*-form* $\tilde{f}$ *on* $\gamma^{-1}(\Delta_r)$ *such that* $\tilde{f}|_{Y_0}=f\wedge\gamma^*(dt)$ *at points of* Y_0 *and* $\int_Y|\tilde{f}|^2e^{-\chi}\le A_r\int_{Y_0}|f|^2e^{-\chi}$.

Locally expressing an element $s^{(m)}$ of $\Gamma(X_0,mK_{X_0})$ as a sum of terms, each of which is the product of a local holomorphic function $s^{(1)}$ and an element $s^{(m-1)}$ of $\Gamma(X_0,(m-1)K_{X_0}+E)$ is precisely Proposition 2.3, necessitating the use of E.

One constructs a metric of $(m-1)K_X+E$ by using the sum of absolute-value squares of elements of $\Gamma(X,(m-1)K_X+E)$ whose restrictions to X_0 form a basis of $\Gamma(X_0,(m-1)K_{X_0}+E)$. Proposition 2.4 is now applicable to show the surjectivity of $\Gamma(X,(m-1)K_X+E)\to\Gamma(X_0,(m-1)K_{X_0}+E)$ by induction on m. To get rid of E, for a sufficiently large ℓ one takes the ℓ-th power of an element of $\Gamma(X_0,mK_{X_0})$ and multiplies it by an element of $\Gamma(X,E)$ and then takes the ℓ-th root of the absolute value after its extension. This process, together with Hölder's inequality, is used to produce a metric of $(m-1)K_X$ which we can use in the application of Proposition 2.4 to get the surjectivity of $\Gamma(X,mK_X)\to\Gamma(X_0,mK_{X_0})$. The assumption of general type facilitates the last technical step of getting rid of E by writing $aK_X=E+D$ for some sufficiently large integer a and an effective divisor D.

Kawamata [6] translated the argument of [17] to a purely algebraic geometric setting and Nakayama [12] explored generalizations including results concerning $\lim_{m\to\infty}\frac{1}{m}\log\dim_{\mathbf{C}}\Gamma(X_t,mK_{X_t}+E)$ as a function of t. The case of non general type necessitates letting ℓ, which is used in taking the power and the root, go to infinity. One has to control the estimates in the limiting process.

Tsuji put on the web a preprint on the deformational invariance of the plurigenera for manifolds not necessarily of general type [26], in which, besides the techniques of [17], he uses his theory of analytic Zariski decomposition and generalized Bergman kernels. Tsuji's approach of generalized Bergman kernels naturally and elegantly reduces the problem of the deformational invariance of the plurigenera to a growth estimate on the generalized Bergman kernels. Unfortunately this crucial estimate is lacking and seems unlikely to be establishable, as explained in [20].

In [20] a metric as singular as possible is introduced for the limiting process, which, together with an estimation technique using the concavity of the logarithmic function, successfully removes the technical assumption of general type in [17].

The deformational invariance of the plurigenera for Kähler manifolds is still open. Only known results on the Kähler case are due to Levine's [9] with the assumption of some pluricanonical section with nonsingular divisor (or only mild singularities). To generalize the methods of [17] and [20] to the Kähler case, one

possibility is to use the "absolute value" of a holomorphic line bundle constructed from the Kähler metric, because in the key argument only the absolute value of the constructed holomorphic section is used and not the section itself. There is still no method of implementing this possibility.

3. Nonexistence of smooth Levi-Flat hypersurface in $\mathbf{P}_2$

The problem of the nonexistence of smooth Levi-flat hypersurface in $\mathbf{P}_2$ has its origin in dynamical systems in $\mathbf{P}_2$ (see [8]). In terms of $\bar\partial$ estimates, its significance is that it gives a natural geometric setting for the understanding of $\bar\partial$ regularity for domains with Levi-flat boundary. The $\bar\partial$ regularity problem for a relatively compact domain Ω with smooth boundary $\partial\Omega$ in a complex manifold is to find a solution u on Ω, smooth up to $\partial\Omega$, to the equation $\bar\partial u = g$ with a given $\bar\partial$-closed $(0,1)$-form g on Ω, smooth up to of $\partial\Omega$. *Global regularity* is said to hold for Ω if regularity holds for the particular solution u, known as the *Kohn solution*, which is orthogonal to all the L^2 holomorphic functions on Ω. The problem of global regularity has been very extensively studied in the past couple of decades (see bibliographies in [2],[7]). Global regularity holds for strictly pseudoconvex domains and, more generally, for weakly pseudoconvex domains whose boundary points are all of finite type. Finite type means that local complex-analytic curves touch the boundary only to bounded finite (normalized) order. Global regularity holds also for weakly pseudoconvex domains defined by global smooth weakly plurisubharmonic functions. On the other hand, worm domains are counter-examples for global regularity for general weakly pseudoconvex domains [2]. Though the nonexistence of smooth Levi-flat hypersurface in $\mathbf{P}_2$ is connected with the regularity of any one solution of the $\bar\partial$-equation rather than the particular Kohn solution, its proof ushers in a new approach of using vector fields to obtain $\bar\partial$ regularity for domains with Levi-flat boundary. The following solution of the Levi-flat hypersurface problem was given in [21].

Theorem 3.1 [21]. *Let $q \geq 8$. Then there exists no C^q Levi-flat real hypersurface M in $\mathbf{P}_2$.*

The nonexistence of real-analytic Levi-flat hypersurface in $\mathbf{P}_3$ was proved by Lins-Neto [8]. Ohsawa [14] treated the nonexistence of real-analytic Levi-flat hypersurface in $\mathbf{P}_2$ (some points in the argument there not yet complete). The nonexistence of smooth Levi-flat hypersurface in $\mathbf{P}_3$ was proved in [19]. The real-analytic case is completely different in nature from the smooth case, because the structure is automatically extendible to a neighborhood for the real-analytic case. Nonexistence in $\mathbf{P}_2$ implies nonexistence in $\mathbf{P}_n$ $(n \geq 2)$ by slicing with a generic linear $\mathbf{P}_2$.

The following argument reduces the problem to a $\bar\partial$ regularity question. Suppose M exists. We seek a contradiction from the positivity of the $(1,0)$-normal bundle $N^{(1,0)}_{M,\mathbf{P}_2}$ of the Levi-flat hypersurface M. The curvature θ of $N^{(1,0)}_{M,\mathbf{P}_2}$ with the metric induced from the Fubini-Study metric is positive, because a quotient bundle cannot be less positive. On the other hand, M is the zero-set of a smooth $\mathbf{R}$-valued function f_M on $\mathbf{P}_2$ with df_M nowhere zero on M. Evaluation by ∂f_M shows that

$N^{(1,0)}_{M,\mathbf{P}_2}$ is smoothly trivial and θ must be d-exact on M, which means that $\theta = d\alpha$ for some smooth real 1-form α on M. Decompose $\alpha = \alpha^{(1,0)} + \alpha^{(0,1)}$ into its $(1,0)$ and $(0,1)$ components. If $\alpha^{(0,1)} = \bar\partial_b \psi$ for some smooth function ψ on M, then $\theta = \sqrt{-1}\partial_b\bar\partial_b\,(2\mathrm{Im}\,\psi)$. At a point of M where $\mathrm{Im}\,\psi$ assumes its maximum, the positivity of θ along the holomorphic foliation is contradicted. The problem is thus reduced to solving the $\bar\partial_b$ equation on M with regularity. By applying the Mayer-Vietoris sequence to $\mathbf{P}_2 - M = U_1 \cup U_2$ and using the vanishing of $H^2(\mathbf{P}_2, \mathcal{O}_{\mathbf{P}_2})$, the problem is reduced to whether, for any $\bar\partial$-closed $(0,1)$-form g on U_j smooth up to ∂U_j, the equation $\bar\partial u = g$ can be solved on U_j with u smooth up to ∂U_j.

The usual approach to $\bar\partial$ regularity is to use the Bochner-Kodaira formula with boundary $\|\bar\partial g\|^2_\Omega + \|\bar\partial^* g\|^2_\Omega = \int_{\partial\Omega} \langle \mathcal{L}, \bar g \wedge g\rangle + \|\bar\nabla g\|^2_\Omega + (\Theta_E, \bar g \wedge g)_\Omega$ (with g being a smooth E-valued $(n,1)$-form in the domain of $\bar\partial^*$), to solve the equation with L^2 estimates and then apply differential operators, integration by parts, and commutation relations to prove regularity. Here $n = \dim_{\mathbf{C}}\Omega$, $\|\cdot\|_\Omega$ is the L^2 norm over Ω, $\bar\partial^*$ is the adjoint of $\bar\partial$, $\bar\nabla$ means covariant differentiation in the $(0,1)$-direction, $\mathcal{L}$ is the Levi form of $\partial\Omega$, and Θ_E is the curvature form of the Hermitian holomorphic line bundle E (which is usually chosen to be trivial).

In our new approach to get $\bar\partial$ regularity for the Levi-flat domain Ω in $\mathbf{P}_2$ the use of holomorphic vector fields compensates for the complete lack of strict positivity for the Levi form of the boundary. We use a new norm to derive the Bochner-Kodaira formula with boundary. We choose a vector field ξ on $\mathbf{P}_2$ which generates biholomorphisms preserving the Fubini-Study metric. The new norm is the L^2 norm $L^2_m(\Omega,\xi)$ for Lie derivatives $(\mathcal{L}ie_\xi)^j\, g$ along ξ for order $j \le m$ on Ω for $(0,1)$-form g. Since ξ generates metric-preserving biholomorphisms, the formal adjoint of $\bar\partial$ with respect to $L^2_m(\Omega,\xi)$ agrees with the one with respect to usual L^2. One usual difficulty with regularity is the error terms from the commutation of differential operators with $\bar\partial$ and $\bar\partial^*$. One advantage of using $\mathcal{L}ie_\xi$ is that there are no error terms from its commutation with $\bar\partial$ and $\bar\partial^*$.

There are two technical problems. One is how to establish, for such a norm, the Bochner-Kodaira formula with boundary. The other is that appropriate regularity for a solution of the $\bar\partial$ equation with finite $L^2_m(\Omega,\xi)$ norm can be obtained only at points where the real and imaginary parts of ξ are not both tangential to $\partial\Omega$. We handle the first problem as follows. We prove that, if g belongs to the domain of the adjoint of $\bar\partial$ with respect to $L^2_m(\Omega,\xi)$, then $(\mathcal{L}ie_\xi)^j\, g$ belongs to the domain of the adjoint of $\bar\partial$ with respect to the usual L^2 norm on Ω for $j \le m$. The formula for the new norm is simply the sum, over $0 \le j \le m$, of such a formula for the usual L^2 norm on Ω for $(\mathcal{L}ie_\xi)^j\, g$. The proof for $(\mathcal{L}ie_\xi)^j\, g$ to belong to the domain of the adjoint of $\bar\partial$ with respect to the usual L^2 norm on Ω consists of two steps. One shows that this is locally true at points where ξ is not tangential to $\partial\Omega$. Then one uses a removable singularity argument to handle the other points when ξ has been chosen generic enough. For the second problem, to handle the other points for a generic ξ, we use the foliation of $\partial\Omega$ by local complex-analytic curves and the generalized Cauchy integral formula along the local complex-analytic curves.

4. Hyperbolicity of generic hypersurface of high degree in $\mathbf{P}_n$

A complex manifold X is *hyperbolic* if there exists no nonconstant holomorphic map $\mathbf{C} \to X$. For the last few decades the study of hyperbolicity has been focussed on hypersurfaces and their complements in two important settings: (1) inside an abelian variety and (2) inside $\mathbf{P}_n$. In the general setting hyperbolicity is conjectured to be linked to the positivity of canonical line bundle in the following formulation.

Conjecture 4.1 *(Conjecture of Green-Griffiths). In a compact algebraic manifold X of general type (or with positive canonical line bundle) there exists a proper subvariety Y containing the images of all nonconstant holomorphic maps $\mathbf{C} \to X$.*

The theory for the setting inside an abelian variety is very well developed (see [16],[18] for references). The Zariski closure of any holomorphic map from $\mathbf{C}$ to an abelian variety A is the translate of an abelian subvariety of A. In particular, a subvariety of an abelian variety A which does not contain any translate of an abelian subvariety of A is hyperbolic. The defect of an ample divisor in an abelian variety is zero. In particular, the complement of an ample divisor in an abelian variety is hyperbolic.

Except those motivated by methods of number theory due to McQuillan, practically all the major techniques for problems related to hyperbolicity in the setting of abelian varieties are due to Bloch [1] who introduced the use of holomorphic jet differentials and differential equations in conjunction with the jet differentials. Investigations on problems related to hyperbolicity in the setting of abelian varieties have essentially been completed. Only technical details such as getting an optimal lower bound for k_n in Theorem 4.2 below remain open. Theorem 4.2 (proved in Addendum of [24]) was added to [24] in response to a difficulty in the proof of Lemma 2 of the original paper [24] pointed out in [13]. The difficulty resulted from an attempt to use semi-continuity of cohomology groups in deformations to avoid employing Bloch's technique from [1] which involves the uniqueness part of the fundamental theorem of ordinary differential equations. Putting back Bloch's technique removes the difficulty and at the same time improves the zero defect statement in [24] to Theorem 4.2 on the second main theorem with truncated multiplicity.

Theorem 4.2 (Addendum, [24]). *Let D be an ample divisor of an abelian variety A of complex dimension n and let $k_0 = 0$, $k_1 = 1$, and $k_{\ell+1} = k_\ell + 3^{n-\ell-1}(4(k+1))^\ell D^n$ $(1 \le \ell < n)$. Then for any holomorphic map $f : \mathbf{C} \to A$ whose image is not contained in any translate of D, the following second main theorem with truncated multiplicity holds: $m(r,f,D) + (N(r,f,D) - N_{k_n}(r,f,D)) = O(\log T(r,f,D) + \log r)$ for r outside some set whose measure with respect to $\frac{dr}{r}$ is finite. Here $T(r,f,D)$, $m(r,f,D)$, $N(r,f,D)$, $N_{k_n}(r,f,D)$ are respectively the characteristic, proximity, counting functions, and truncated counting functions.*

For the setting inside $\mathbf{P}_n$ there is the following outstanding conjecture.

Conjecture 4.3 (Kobayashi's Conjecture). (a) *The complement in $\mathbf{P}_n$ of a generic hypersurface of degree at least $2n+1$ is hyperbolic.* (b) *A generic hypersurface of degree at least $2n-1$ in $\mathbf{P}_n$ is hyperbolic for $n \ge 3$.*

For Conjecture 4.3(a) the complement in $\mathbf{P}_2$ of a generic curve of sufficiently

high degree is known to be hyperbolic [23]. For Conjecture 4.3(b) a generic surface of degree ≥ 36 in $\mathbf{P}_3$ is known to be hyperbolic [10]. The degree bound is lowered to 21 in [4]. There are some constructions of examples of smooth hyperbolic hypersurfaces in $\mathbf{P}_n$ (see [15]). The hyperbolicity result we want to discuss here is the following.

Theorem 4.4 [22]. *There exists a positive integer δ_n such that a generic hypersurface in $\mathbf{P}_n$ of degree $\geq \delta_n$ is hyperbolic.*

We sketch its proof here. A central role will be played by jet differentials which we now define. A k-jet differential on a complex manifold X with local coordinates $x_1, \cdots, x_n$ is locally a polynomial in $d^\ell x_j$ $(1 \leq \ell \leq k, 1 \leq j \leq n)$.

Lemma 4.5 (Lemma of Jet Differentials). *If a holomorphic jet differential ω on a compact complex manifold X vanishes on an ample divisor of X and $\varphi : \mathbf{C} \to X$ is a holomorphic map, then $\varphi^*\omega$ is identically zero on $\mathbf{C}$.*

The intuitive reason for Lemma 4.5 is that $\mathbf{C}$ does not admit a metric (or not even a *k-jet metric*) with curvature bounded above by negative number. While a usual metric assigns a value to a tangent vector (which is a 1-jet), a k-jet metric assigns a value to a k-jet. A non identically zero $\varphi^*\omega$ defines a k-jet metric $|\varphi^*\omega|^2$ on $\mathbf{C}$ which, even with some degeneracy, still gives a contradiction by its negative curvature. A rigorous proof of Lemma 4.5 depends on the logarithmic derivative lemma of Nevanlinna theory. A consequence of Lemma 4.5 is that the image of the k-jet $d^k\varphi$ of any holomorphic map $\varphi : \mathbf{C} \to X$ satisfies the differential equation $\omega = 0$ on X. If there exist enough independent such ω on X, then the system of all equations $\omega = 0$ does not admit any local solution curve and X is hyperbolic.

In the setting of abelian varieties Bloch constructed jet differentials by comparing meromorphic functions on the image and the target of a map with finite fibers. For a holomorphic map φ from $\mathbf{C}$ to an abelian variety A, let X be the Zariski closure of the image of φ in A and $\mathcal{X}$ be the Zariski closure of $(d^k\varphi)(\mathbf{C})$ in $J_k(A) = A \times \mathbf{C}^{nk}$. Here $J_k(\cdot)$ means the space of k-jets. Let $\sigma_k : \mathcal{X} \to \mathbf{C}^{nk}$ be induced by the natural projection $J_k(A) = A \times \mathbf{C}^{nk} \to \mathbf{C}^{nk}$ which forgets the position and keeps the differentials. Let $\tau : J_k(X) \to X$ be the natural projection. Let F be a meromorphic function on X whose pole-set is some ample divisor D. Suppose $\sigma_k : \mathcal{X} \to \mathbf{C}^{nk}$ is generically finite. Let $x_1, \cdots, x_n$ be the coordinates of $\mathbf{C}^n$. Then $F \circ \tau$ belongs to a finite extension of the rational function field of $\mathbf{C}^{nk}$ and there exist polynomials P_j $(0 \leq j \leq p)$ with constant coefficients in the variables $d^\ell x_\nu$ $(1 \leq \ell \leq k, 1 \leq \nu \leq n)$ such that $\sum_{j=0}^{p} (\sigma_k^* P_j)(\tau^* F)^j = 0$ on $\mathcal{X}$ and $\sigma_k^* P_p$ is not identically zero on $\mathcal{X}$. The equation forces the holomorphic jet differential P_p on $\mathcal{X}$ to vanish on the ample divisor $\tau^{-1}(D)$. The assumption of generical finiteness of σ_k is tied to the translational invariance of X.

The idea of our method of construction of holomorphic jet differentials on a generic hypersurface X in $\mathbf{P}_n$ defined of by a polynomial f of degree δ is to use the theorem of Riemann-Roch and the lower bound of negativity of jet differential bundles of X. The theorem of Riemann-Roch was first used by Green-Griffiths to obtain holomorphic jet differentials and is applicable only for surfaces where the higher cohomology groups could be easily handled. We can handle the higher cohomology groups in our higher dimensional case because of the lower bound of negativity of jet differential bundles of X. Since the twisted cohomology groups

of $\mathbf{P}_n$ comes from counting the number of monomials, in the actual proof direct counting of monomials is used. Let $x_1, \cdots, x_n$ (respectively $z_0, \cdots, z_n$) be the inhomogeneous (respectively homogeneous) coordinates of $\mathbf{P}_n$. Let Q be a non identically zero polynomial of degree m_0 in $x_1, \cdots, x_n$ and of homogeneous weight m in $d^j x_\ell$ $(1 \le j \le n-1, 1 \le \ell \le n)$ with the weight of $d^j x_\ell$ equal to j. If $m_0 + 2m < \delta$, then Q is not identically zero on $J_{n-1}(X)$. By counting the number of coefficients of Q and the number of equations needed for the jet differential on X defined by Q to vanish on an ample divisor in X of high degree defined by a polynomial $g = 0$ in $\mathbf{P}_n$ and using $f = df = \cdots = d^{n-1}f = 0$ to eliminate one coordinate and its differentials, we obtain a jet differential $\frac{Q}{g}$ on X which is holomorphic and vanishes on an ample divisor of high degree.

Proposition 4.6 (Existence of Holomorphic Jet Differentials). *If $0 < \theta_0, \theta, \theta' < 1-\epsilon$ with $n\theta_0 + \theta \ge n + \epsilon$, then there exists an explicit $A = A(n, \epsilon) > 0$ such that for $\delta \ge A$ there exists a non identically zero $\mathcal{O}_{\mathbf{P}_n}(-q)$-valued holomorphic $(n-1)$-jet differential ω on X of total weight m with $q \ge \delta^{\theta'}$ and $m \le \delta^{\theta}$.*

To construct enough independent jet differentials, we use meromorphic vector fields of low pole order on the total space $\mathcal{X}$ of all hypersurfaces in $\mathbf{P}_n$ of degree δ. The total space $\mathcal{X}$ is defined by $f = \sum_{\nu \in \mathbf{N}^{n+1}, |\nu| = \delta} \alpha_\nu z^\nu$ of bidegree $(\delta, 1)$ in $\mathbf{P}_n \times \mathbf{P}_N$, where $N = \binom{\delta+n}{n} - 1$, $z^\nu = z_0^{\nu_0} \cdots z_n^{\nu_n}$, $|\nu| = \nu_0 + \cdots + \nu_n$, and $\mathbf{N}$ is the set of all nonnegative integers. Let $e_\ell = (0, \cdots, 0, 1, \cdots, 0) \in \mathbf{N}^{n+1}$ with 1 in the ℓ-th place. The $(1,0)$-twisted tangent bundle of $\mathcal{X}$ is globally generated by holomorphic sections of the forms $L\left(z_q \left(\frac{\partial}{\partial \alpha_{\lambda + e_p}}\right) - z_p \left(\frac{\partial}{\partial \alpha_{\lambda + e_q}}\right)\right)$ and $\sum_j B_j \frac{\partial}{\partial z_j} + \sum_\mu L_\mu \frac{\partial}{\partial \alpha_\mu}$, where $\lambda \in \mathbf{N}^{n+1}$ with $|\lambda| = \delta - 1$ and L, L_μ (respectively B_j) are homogeneous linear functions of $\{\alpha_\nu\}$ (respectively $z_0, \cdots, z_n$) with L_μ and B_j suitably chosen.

We introduce the space $J_{n-1}^{\text{vert}}(\mathcal{X})$ of vertical $(n-1)$-jets of $\mathcal{X}$ which is defined by $f = df = \cdots = d^{n-1}f = 0$ in $(J_{n-1}(\mathbf{P}_n)) \times \mathbf{P}_N$ with the coefficients α_ν of f regarded as constants when forming $d^j f$. By generalizing the above construction of vector fields on $\mathcal{X}$ to vector fields on $J_{n-1}^{\text{vert}}(\mathcal{X})$, one obtains the following.

Proposition 4.7 (Existence of Low Pole-Order Vector Fields). *There exist $c_n, c_n' \in \mathbf{N}$ such that the (c_n, c_n')-twisted tangent bundle of the projectivization of $J_{n-1}^{\text{vert}}(\mathcal{X})$ is globally generated. (To avoid considering the singularities of weighted projective spaces, one can interpret the statement by using functions which are polynomials of homogeneous weight along the fibers of $J_{n-1}^{\text{vert}}(\mathcal{X})$.)*

For a generic fiber X of $\mathcal{X}$ the constructed holomorphic $(n-1)$-jet differential ω on X with vanishing order at least q on the infinity divisor can be extended holomorphically to $\tilde{\omega}$ on all neighboring fibers with vanishing order at least q on the infinity divisor. We use vector fields $v_1, \cdots, v_p$ on $J_{n-1}^{\text{vert}}(\mathcal{X})$ with fiber pole order low relative to q and take successive Lie derivatives $\mathcal{L}ie_{v_1} \cdots \mathcal{L}ie_{v_p} \tilde{\omega}$ whose restrictions to X give holomorphic jet differentials on X vanishing on an ample divisor of X. Because of the bound on the weight m in the construction of ω, for δ sufficiently large the jet differentials from the Lie derivatives are independent enough to eliminate the derivatives from the differential equations they define. As a result, one concludes that for some proper subvariety Y in X the image of any nonconstant holomorphic map from $\mathbf{C}$ to X is contained in Y.

To get the full conclusion of hyperbolicity, for the constructed ω one has to

control the vanishing order of the coefficients of the monomials of the differentials. For a generic X the construction process enables one to bound the vanishing order by $\delta^{2-\eta}$ for some $\eta > 0$. For hyperbolicity one needs the better bound of $\delta^{1-\eta}$ for some $\eta > 0$. To achieve it, one uses an appropriate embedding $\Phi : \mathbf{P}_n \to \mathbf{P}_{\hat{n}}$ of degree δ_1 so that a generic hypersurface X of degree $\delta := \delta_1\delta_2$ in $\mathbf{P}_n$ can be extended to a hypersurface $\hat{X}$ of degree δ_2 in $\mathbf{P}_{\hat{n}}$. For this step the method of multiplier ideal sheaves from $\bar{\partial}$ estimates is used. We deform Φ slightly and pull back the jet differential $\hat{\omega}$ constructed on $\hat{X}$ to get a differential ω on a slight deformation of X. When the image of the deformed Φ has appropriate transversality to the zero set of the coefficients of $\hat{\omega}$, an appropriate choice of δ_1 and δ_2 gives the required bound on the vanishing order of the coefficients of ω. This is at the expense of increasing the order of the jet differential from $n-1$ to $\hat{n}-1$, which does not affect the argument. For this additional step in the argument the degree δ must be a product. To remove this condition, one uses an embedding $\mathbf{P}_n \to \mathbf{P}_{\hat{n}_1} \times \mathbf{P}_{\hat{n}_2}$ instead of Φ.

The use and the construction of meromorphic vector fields on $J^{\mathrm{vert}}_{n-1}(\mathcal{X})$ of low pole order along the fibers are motivated by Clemens's work [3] (with later generalizations and improvements by Ein [5] and Voisin [25]) on the nonexistence of regular rational and elliptic curves on generic hypersufaces of sufficiently high degree.

There is no way yet to handle Conjecture 4.1. Additional assumptions such as $K_X - mL > 0$ or $(K_X - mL)\, L^{n-1} > 0$ for some large m and L ample or very ample could facilitate the construction of holomorphic jet differentials vanishing on an ample divisor. One possibility to handle the question of enough independent jet differentials is to deform $d^k\varphi$ of φ for each $k \geq 1$ separately and use techniques analogous to the twisted difference maps in the Vojta-Faltings proof of the Mordell conjecture and to McQuillan's separate rescaling of an entire holomorphic curve in each factor of a product of several copies of an abelian variety (see pp.504-505,[16]).

References

[1] A. Bloch, Sur les systèmes de fonctions uniformes satisfaisant à l'équation d'une variété algébrique dont l'irrégularité dépasse la dimension, *J. de Math.* 5 (1926), 19–66.

[2] M. Christ, Global C^∞ irregularity of the $\bar{\partial}$-Neumann problem for worm domains, *J. of Amer. Math. Soc.* 9 (1996), 1171–1185.

[3] H. Clemens, Curves on generic hypersurfaces, *Ann. Ec. Norm. Sup.* 19 (1986), 629–636.

[4] J.-P. Demailly & J. El Goul, Hyperbolicity of generic surfaces of high degree in projective 3-space, *Amer. J. Math.* 122 (2000), 515–546.

[5] L. Ein, Subvarieties of generic complete intersections, *Invent. Math.* 94 (1988), 163–169. II. *Math. Ann.* 289 (1991), 465–471.

[6] Y. Kawamata, *Deformations of canonical singularities. J. Amer. Math. Soc.* 12 (1999), no. 1, 85–92.

[7] J. J. Kohn, Subellipticity of the $\bar{\partial}$-Neumann problem on pseudo-convex domains: sufficient conditions. *Acta Math.* 142 (1979), 79–122.

[8] A. Lins-Neto, A note on projective Levi-flats and minimal sets of algebraic functions. *Ann. Inst. Fourier* 49 (1999), 1369–1385.
[9] M. Levine, Pluri-canonical divisors on Kähler manifolds, *Invent. math.* 74 (1983), 293–903. II. *Duke Math. J.* 52 (1985), no. 1, 61–65.
[10] M. McQuillan, Diophantine approximations and foliations. *Inst. Hautes Études Sci. Publ. Math.* 87 (1998), 121–174.
[11] A. Nadel, Multiplier ideal sheaves and the existence of Kähler-Einstein metrics of positive scalar curvature, *Ann. of Math.*, 132 (1989), 549–596.
[12] N. Nakayama, Invariance of the plurigenera of algebraic varieties. *Research Inst. for Math. Sci., RIMS*, No. 1191, Preprint, March 1998.
[13] J. Noguchi, J. Winkelmann, & K. Yamanoi, The second main theorem for holomorphic curves into semi-Abelian varieties, Preprint, Univ. Tokyo, 1999.
[14] T. Ohsawa, Nonexistence of real analytic Levi flat hypersurfaces in $\mathbf{P}^2$, *Nagoya Math. J.* 158 (2000), 95–98.
[15] B. Shiffman & M. Zaidenberg, Hyperbolic hypersurfaces in $\mathbf{P}_n$ of Fermat-Waring type. *Proc. Amer. Math. Soc.* 130 (2002), 2031–2035.
[16] Y.-T. Siu, Hyperbolicity problems in function theory. In: *Five Decades as a Mathematician and Educator*, ed. K.-Y.Chan & M.-C. Liu, World Scientific 1995, pp. 409–513.
[17] Y.-T. Siu, Invariance of plurigenera. *Invent. Math.* 134 (1998), 661–673.
[18] Y.-T. Siu, Recent techniques in hyperbolicity problems. In: *Several complex variables*, ed. M. Schneider & Y.-T. Siu, Cambridge University Press, Cambridge, 1999, pp. 429–508.
[19] Y.-T. Siu, Nonexistence of smooth Levi-flat hypersurfaces in complex projective spaces of dimension ≥ 3, *Ann. of Math.* 151 (2000), 1217–1243.
[20] Y.-T. Siu, Extension of twisted pluricanonical sections with plurisubharmonic weight and invariance of semipositively twisted plurigenera for manifolds not necessarily of general type. In: *Complex Geometry*, ed. I. Bauer et al, Springer-Verlag 2002, pp. 223–277.
[21] Y.-T. Siu, $\bar{\partial}$-regularity for weakly pseudoconvex domains in compact Hermitian symmetric spaces with respect to invariant metrics. *Ann. of Math.* 157 (2002), 1–27.
[22] Y.-T. Siu, Hyperbolicity of generic high-degree hypersurfaces in complex projective spaces. Preprint 2002.
[23] Y.-T. Siu & S.-K. Yeung, Hyperbolicity of the complement of a generic smooth curve of high degree in the complex projective plane. *Invent. Math.* 124 (1996), 573–618.
[24] Y.-T. Siu & S.-K. Yeung, Defects for ample divisors of abelian varieties, Schwarz lemma, and hyperbolic hypersurfaces of low degrees. *Amer. J. Math.* 119 (1997), no. 5, 1139–1172. Addendum, Preprint 2000.
[25] C. Voisin, On a conjecture of Clemens on rational curves on hypersurfaces, *J. Diff. Geom.* 44 (1996), 200–213.
[26] H. Tsuji, Deformational invariance of plurigenera, Preprint 2001. math.AG/0012225. (Earlier version: Invariance of plurigenera of varieties with nonnegative Kodaira dimensions. math.AG/0011257.)

ICM 2002 · Vol. I · 449–474

Galois Representations*

R. Taylor†

Abstract

In the first part of this paper we try to explain to a general mathematical audience some of the remarkable web of conjectures linking representations of Galois groups with algebraic geometry, complex analysis and discrete subgroups of Lie groups. In the second part we briefly review some limited recent progress on these conjectures.

2000 Mathematics Subject Classification: 11F80.
Keywords and Phrases: Galois representations, L-function, Automorphic forms.

Introduction

The organisers requested a talk which would both be a colloquium style talk understandable to a wide spectrum of mathematicians and one which would survey the recent developments in the subject. I have found it hard to meet both desiderata, and have opted to concentrate on the former. Thus the first three sections of this paper contain a simple presentation of a web of deep conjectures connecting Galois representations to algebraic geometry, complex analysis and discrete subgroups of Lie groups. This will be of no interest to the specialist. My hope is that the result is not too banal and that it will give the non-specialist some idea of what motivates work in this area. I should stress that nothing I write here is original. In the final section I briefly review some of what is known about these conjectures and *very briefly* mention some of the available techniques. I also mention two questions which lie outside the topic we are discussing, but which would have important implications for it. Maybe someone can make progress on them?

Due to lack of space much of this article is too abbreviated. A somewhat expanded version is available on my website `www.math.harvard.edu/~rtaylor` and will hopefully be published elsewhere.

*The work on this article was partially supported by NSF Grant DMS-9702885.
†Department of Mathematics, Harvard University, 1 Oxford St., Cambridge, MA 02138, USA. E-mail: rtaylor@math.harvard.edu

1. Galois representations

We will let $\mathbb{Q}$ denote the field of rational numbers and $\overline{\mathbb{Q}}$ denote the field of algebraic numbers, the algebraic closure of $\mathbb{Q}$. We will also let $G_{\mathbb{Q}}$ denote the group of automorphisms of $\overline{\mathbb{Q}}$, that is $\mathrm{Gal}\,(\overline{\mathbb{Q}}/\mathbb{Q})$, the absolute Galois group of $\mathbb{Q}$. Although it is not the simplest it is arguably the most natural Galois group to study. An important technical point is that $G_{\mathbb{Q}}$ is naturally a (profinite) topological group, a basis of open neighbourhoods of the identity being given by the subgroups $\mathrm{Gal}\,(\overline{\mathbb{Q}}/K)$ as K runs over subextensions of $\overline{\mathbb{Q}}/\mathbb{Q}$ which are finite over $\mathbb{Q}$.

To my mind the Galois theory of $\mathbb{Q}$ is most interesting when one looks not only at $G_{\mathbb{Q}}$ as an abstract (topological) group, but as a group with certain additional structures associated to the prime numbers. I will now briefly describe these structures.

For each prime number p we may define an absolute value $|\ |_p$ on $\mathbb{Q}$ by setting

$$|\alpha|_p = p^{-r}$$

if $\alpha = p^r a/b$ with a and b integers coprime to p. If we complete $\mathbb{Q}$ with respect to this absolute value we obtain the field of p-adic numbers $\mathbb{Q}_p$, a totally disconnected, locally compact topological field. We will write $G_{\mathbb{Q}_p}$ for its absolute Galois group, $\mathrm{Gal}\,(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$. The absolute value $|\ |_p$ has a unique extension to an absolute value on $\overline{\mathbb{Q}}_p$ and $G_{\mathbb{Q}_p}$ is identified with the group of automorphisms of $\overline{\mathbb{Q}}_p$ which preserve $|\ |_p$, or equivalently the group of continuous automorphisms of $\overline{\mathbb{Q}}_p$. For each embedding $\overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_p$ we obtain a closed embedding $G_{\mathbb{Q}_p} \hookrightarrow G_{\mathbb{Q}}$ and as the embedding $\overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_p$ varies we obtain a conjugacy class of closed embeddings $G_{\mathbb{Q}_p} \hookrightarrow G_{\mathbb{Q}}$. Slightly abusively, we shall consider $G_{\mathbb{Q}_p}$ a closed subgroup of $G_{\mathbb{Q}}$, suppressing the fact that the embedding is only determined up to conjugacy.

This can be compared with the situation 'at infinity'. Let $|\ |_\infty$ denote the usual Archimedean absolute value on $\mathbb{Q}$. The completion of $\mathbb{Q}$ with respect to $|\ |_\infty$ is the field of real numbers $\mathbb{R}$ and its algebraic closure is $\mathbb{C}$ the field of complex numbers. Each embedding $\overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$ gives rise to a closed embedding

$$\{1, c\} = G_{\mathbb{R}} = \mathrm{Gal}\,(\mathbb{C}/\mathbb{R}) \hookrightarrow G_{\mathbb{Q}}.$$

As the embedding $\overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$ varies one obtains a conjugacy class of elements $c \in G_{\mathbb{Q}}$ of order 2, which we refer to as complex conjugations.

There are however many important differences between the case of finite places (i.e. primes) and the infinite place $|\ |_\infty$. For instance $\overline{\mathbb{Q}}_p/\mathbb{Q}_p$ is an infinite extension and $\overline{\mathbb{Q}}_p$ is not complete. We will denote its completion by $\mathbb{C}_p$. The Galois group $G_{\mathbb{Q}_p}$ acts on $\mathbb{C}_p$ and is in fact the group of continuous automorphisms of $\mathbb{C}_p$.

The elements of $\mathbb{Q}_p$ (resp. $\overline{\mathbb{Q}}_p$) with absolute value less than or equal to 1, form a closed subring $\mathbb{Z}_p$ (resp. $\mathcal{O}_{\overline{\mathbb{Q}}_p}$). These rings are local with maximal ideals $p\mathbb{Z}_p$ (resp. $\mathfrak{m}_{\overline{\mathbb{Q}}_p}$) consisting of the elements with absolute value strictly less than 1. The field $\mathcal{O}_{\overline{\mathbb{Q}}_p}/\mathfrak{m}_{\overline{\mathbb{Q}}_p}$ is an algebraic closure of the finite field with p elements $\mathbb{F}_p = \mathbb{Z}_p/p\mathbb{Z}_p$, and we will denote it by $\overline{\mathbb{F}}_p$. Thus we obtain a continuous map

$$G_{\mathbb{Q}_p} \longrightarrow G_{\mathbb{F}_p}$$

which is surjective. Its kernel is called the inertia subgroup of $G_{\mathbb{Q}_p}$ and is denoted by $I_{\mathbb{Q}_p}$. The group $G_{\mathbb{F}_p}$ is procyclic and has a canonical generator called the (geometric) Frobenius element and defined by

$$\mathrm{Frob}_p^{-1}(x) = x^p.$$

In many circumstances it is technically convenient to replace $G_{\mathbb{Q}_p}$ by a dense subgroup $W_{\mathbb{Q}_p}$, which is referred to as the Weil group of $\mathbb{Q}_p$ and which is defined as the subgroup of $\sigma \in G_{\mathbb{Q}_p}$ such that σ maps to

$$\mathrm{Frob}_p^{\mathbb{Z}} \subset G_{\mathbb{F}_p}.$$

We endow $W_{\mathbb{Q}_p}$ with a topology by decreeing that $I_{\mathbb{Q}_p}$ with its usual topology should be an open subgroup of $W_{\mathbb{Q}_p}$.

We will take a moment to describe some of the finer structure of $I_{\mathbb{Q}_p}$ which we will need for technical purposes later. First of all there is a (not quite canonical) continuous surjection

$$I_{\mathbb{Q}_p} \twoheadrightarrow \prod_{l \neq p} \mathbb{Z}_l$$

such that

$$t(\mathrm{Frob}_p \sigma \mathrm{Frob}_p^{-1}) = p^{-1} t(\sigma)$$

for all $\sigma \in I_{\mathbb{Q}_p}$. The kernel of t is a pro-p-group called the wild inertia group. The fixed field $\overline{\mathbb{Q}}_p^{\ker t}$ is obtained by adjoining $\sqrt[n]{p}$ to $\overline{\mathbb{Q}}_p^{I_{\mathbb{Q}_p}}$ for all n coprime to p and

$$\sigma \sqrt[n]{p} = \zeta_n^{t(\sigma)} \sqrt[n]{p},$$

for some primitive n^{th}-root of unity ζ_n (independent of σ, but dependent on t).

In my opinion the most interesting question about $G_{\mathbb{Q}}$ is to describe it together with the distinguished subgroups $G_{\mathbb{R}}$, $G_{\mathbb{Q}_p}$, $I_{\mathbb{Q}_p}$ and the distinguished elements $\mathrm{Frob}_p \in G_{\mathbb{Q}_p}/I_{\mathbb{Q}_p}$.

I want to focus here on attempts to describe $G_{\mathbb{Q}}$ via its representations. Perhaps the most obvious representations to consider are those representations

$$G_{\mathbb{Q}} \longrightarrow GL_n(\mathbb{C})$$

with open kernel, and these so called Artin representations are already very interesting. However one obtains a richer theory if one considers representations

$$G_{\mathbb{Q}} \longrightarrow GL_n(\overline{\mathbb{Q}}_l)$$

which are continuous with respect to the l-adic topology on $GL_n(\overline{\mathbb{Q}}_l)$. We refer to these as *$l$-adic representations.*

One justification for considering l-adic representations is that they arise naturally from geometry. Here are some examples of l-adic representations.

1. A choice of embeddings $\overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$ and $\overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_l$ establishes a bijection between isomorphism classes of Artin representations and isomorphism classes of l-adic representations with open kernel. Thus Artin representations are a special case of l-adic representations: those with finite image.

2. There is a unique character

$$\chi_l : G_{\mathbb{Q}} \longrightarrow \mathbb{Z}_l^{\times} \subset \overline{\mathbb{Q}}_l^{\times}$$

such that

$$\sigma\zeta = \zeta^{\chi_l(\sigma)}$$

for all l-power roots of unity ζ. This is called the l-adic *cyclotomic character.*

3. If $X/\mathbb{Q}$ is a smooth projective variety (and we choose an embedding $\overline{\mathbb{Q}} \subset \mathbb{C}$) then the natural action of $G_{\mathbb{Q}}$ on the cohomology

$$H^i(X(\mathbb{C}), \overline{\mathbb{Q}}_l) \cong H^i_{\text{et}}(X \times_{\mathbb{Q}} \overline{\mathbb{Q}}, \overline{\mathbb{Q}}_l)$$

is an l-adic representation. For instance if $E/\mathbb{Q}$ is an elliptic curve then we have the concrete description

$$H^1_{\text{et}}(E \times_{\mathbb{Q}} \overline{\mathbb{Q}}, \overline{\mathbb{Q}}_l) \cong \text{Hom}_{\mathbb{Z}_l}(\varprojlim_r E[l^r](\overline{\mathbb{Q}}), \overline{\mathbb{Q}}_l) \cong \overline{\mathbb{Q}}_l^2,$$

where $E[l^r]$ denotes the l^r-torsion points on E. We will write $H^i(X(\mathbb{C}), \overline{\mathbb{Q}}_l(j))$ for the twist

$$H^i(X(\mathbb{C}), \overline{\mathbb{Q}}_l) \otimes \chi_l^j.$$

Before discussing l-adic representations of $G_{\mathbb{Q}}$ further, let us take a moment to look at l-adic representations of $G_{\mathbb{Q}_p}$. The cases $l \neq p$ and $l = p$ are very different. Consider first the much easier case $l \neq p$. Here l-adic representations of $G_{\mathbb{Q}_p}$ are not much different from representations of $W_{\mathbb{Q}_p}$ with open kernel. More precisely define a *WD-representation* of $W_{\mathbb{Q}_p}$ over a field E to be a pair

$$r : W_{\mathbb{Q}_p} \longrightarrow GL(V)$$

and

$$N \in \text{End}\,(V),$$

where V is a finite dimensional E-vector space, r is a representation with open kernel and N is a nilpotent endomorphism which satisfies

$$r(\phi)Nr(\phi^{-1}) = p^{-1}N$$

for every lift $\phi \in W_{\mathbb{Q}_p}$ of Frob_p. The key point here is that there is no reference to a topology on E, indeed no assumption that E is a topological field. Given r there are up to isomorphism only finitely many choices for the pair (r, N) and these can be explicitly listed without difficulty. A WD-representation (r, N) is called *unramified* if $N = 0$ and $r(I_{\mathbb{Q}_p}) = \{1\}$. It is called Frobenius semi-simple if r is semi-simple. Any WD-representation (r, N) has a canonical Frobenius semi-simplification $(r, N)^{\text{ss}}$ (see [Tat]). In the case that $E = \overline{\mathbb{Q}}_l$ we call (r, N) l-integral if all the eigenvalues of $r(\phi)$ have absolute value 1. This is independent of the choice of Frobenius lift ϕ.

If $l \neq p$, then there is an equivalence of categories between l-integral WD-representations of $W_{\mathbb{Q}_p}$ over $\overline{\mathbb{Q}}_l$ and l-adic representations of $G_{\mathbb{Q}_p}$. To describe it choose a Frobenius lift $\phi \in W_{\mathbb{Q}_p}$ and a surjection $t_l : I_{\mathbb{Q}_p} \twoheadrightarrow \mathbb{Z}_l$. Up to natural

isomorphism the equivalence does not depend on these choices. We associate to an l-integral WD-representation (r, N) the unique l-adic representation sending

$$\phi^n \sigma \longmapsto r(\phi^n \sigma) \exp(t_l(\sigma) N)$$

for all $n \in \mathbb{Z}$ and $\sigma \in I_{\mathbb{Q}_p}$. The key point is Grothendieck's observation that for $l \neq p$ any l-adic representation of $G_{\mathbb{Q}_p}$ must be trivial on some open subgroup of the wild inertia group. We will write $\mathrm{WD}_p(R)$ for the WD-representation associated to an l-adic representation R. Note that $\mathrm{WD}_p(R)$ is unramified if and only if $R(I_p) = \{1\}$. In this case we call R *unramified.*

The case $l = p$ is much more complicated because there are many more p-adic representations of $G_{\mathbb{Q}_p}$. These have been extensively studied by Fontaine and his co-workers. They single out certain p-adic representations which they call *de Rham* representations. I will not recall the somewhat involved definition here (see however [Fo2] and [Fo3]), but note that 'most' p-adic representations of $G_{\mathbb{Q}_p}$ are not de Rham. To any de Rham representation R of $G_{\mathbb{Q}_p}$ on a $\overline{\mathbb{Q}}_p$-vector space V they associate the following.

1. A WD-representation $\mathrm{WD}_p(R)$ of $W_{\mathbb{Q}_p}$ over $\overline{\mathbb{Q}}_p$ (see [Berg] and [Fo4]).
2. A multiset $\mathrm{HT}(R)$ of $\dim V$ integers, called the Hodge-Tate numbers of R. The multiplicity of i in $\mathrm{HT}(R)$ is

$$\dim_{\overline{\mathbb{Q}}_p}(V \otimes_{\mathbb{Q}_p} \mathbb{C}_p(i))^{G_{\mathbb{Q}_p}},$$

where $\mathbb{C}_p(i)$ denotes $\mathbb{C}_p$ with $G_{\mathbb{Q}_p}$-action $\chi_p(\sigma)^i$ times the usual (Galois) action on $\mathbb{C}_p$.

We now return to the global situation (i.e. to the study of $G_{\mathbb{Q}}$). The l-adic representations of $G_{\mathbb{Q}}$ that arise 'in nature', by which I mean 'from geometry', have a number of very special properties which I will now list. Let $R : G_{\mathbb{Q}} \longrightarrow GL(V)$ be a subquotient of $H^i(X(\mathbb{C}), \overline{\mathbb{Q}}_l(j))$ for some smooth projective variety $X/\mathbb{Q}$ and some integers $i \geq 0$ and j.

1. (Grothendieck) The representation R is unramified at all but finitely many primes p.
2. (Fontaine, Messing, Faltings, Kato, Tsuji, de Jong, see e.g. [Il], [Bert]) The representation R is de Rham in the sense that its restriction to $G_{\mathbb{Q}_l}$ is de Rham.
3. (Deligne, [De]) The representation R is *pure* of weight $w = i - 2j$ in the following sense. There is a finite set of primes S, such that for $p \notin S$, the representation R is unramified at p and for every eigenvalue α of $R(\mathrm{Frob}_p)$ and every embedding $\iota : \overline{\mathbb{Q}}_l \hookrightarrow \mathbb{C}$

$$|\iota\alpha|_\infty^2 = p^w.$$

In particular α is algebraic (i.e. $\alpha \in \overline{\mathbb{Q}}$).

An amazing conjecture of Fontaine and Mazur (see [Fo1] and [FM]) asserts that any irreducible l-adic representation of $G_{\mathbb{Q}}$ satisfying the first two of these properties arises from geometry in the above sense and so in particular also satisfies the third property.

Conjecture 1.1 (Fontaine-Mazur) *Suppose that*

$$R : G_{\mathbb{Q}} \longrightarrow GL(V)$$

is an irreducible l-adic representation which is unramified at all but finitely many primes and with $R|_{G_{\mathbb{Q}_l}}$ de Rham. Then there is a smooth projective variety $X/\mathbb{Q}$ and integers $i \geq 0$ and j such that V is a subquotient of $H^i(X(\mathbb{C}), \overline{\mathbb{Q}}_l(j))$. In particular R is pure of some weight $w \in \mathbb{Z}$.

We will discuss the evidence for this conjecture later. We will call an l-adic representation satisfying the conclusion of this conjecture *geometric.*

Algebraic geometers have formulated some very precise conjectures about the action of $G_{\mathbb{Q}}$ on the cohomology of varieties. We don't have the space here to discuss these in general, but we will formulate, in an as algebraic a way as possible, some of their conjectures.

Conjecture 1.2 (Tate) *Suppose that $X/\mathbb{Q}$ is a smooth projective variety. Then there is a decomposition*

$$H^i(X(\mathbb{C}), \overline{\mathbb{Q}}) = \bigoplus_j M_j$$

with the following properties.

1. *For each prime l and for each embedding $\iota : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_l$, $M_j \otimes_{\overline{\mathbb{Q}},\iota} \overline{\mathbb{Q}}_l$ is an irreducible subrepresentation of $H^i(X(\mathbb{C}), \overline{\mathbb{Q}}_l)$.*
2. *For all indices j and for all primes p there is a WD-representation $\mathrm{WD}_p(M_j)$ of $W_{\mathbb{Q}_p}$ over $\overline{\mathbb{Q}}$ such that*

$$\mathrm{WD}_p(M_j) \otimes_{\overline{\mathbb{Q}},\iota} \overline{\mathbb{Q}}_l \cong \mathrm{WD}_p(M_j \otimes_{\overline{\mathbb{Q}},\iota} \overline{\mathbb{Q}}_l)$$

   for all primes l and all embeddings $\iota : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_l$.
3. *There is a multiset of integers $\mathrm{HT}(M_j)$ such that*
   (a) for all primes l and all embeddings $\iota : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_l$

$$\mathrm{HT}(M_j \otimes_{\overline{\mathbb{Q}},\iota} \overline{\mathbb{Q}}_l) = HT(M_j)$$

   (b) and for all $\iota : \overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$

$$\dim_{\mathbb{C}}((M_j \otimes_{\overline{\mathbb{Q}},\iota} \mathbb{C}) \cap H^{a,i-a}(X(\mathbb{C}), \mathbb{C}))$$

   is the multiplicity of a in $HT(M_j)$.

If one considers the whole of $H^i(X(\mathbb{C}), \overline{\mathbb{Q}})$ rather than its pieces M_j, then part 2. is known to hold up to Frobenius semisimplification for all but finitely many p and part 3. is known to hold (see [Il]). It follows from a theorem of Faltings [Fa] that the whole conjecture is true for H^1 of an abelian variety. The putative constituents M_j are one incarnation of what people call 'pure motives'.

If one believes conjectures 1.1 and 1.2 then 'geometric' l-adic representations should come in compatible families as l varies. There are many ways to make precise the notion of such a compatible family. Here is one.

By a *weakly compatible system of l-adic representations* $\mathcal{R} = \{R_{l,\iota}\}$ we shall mean a collection of semi-simple l-adic representations

$$R_{l,\iota} : G_{\mathbb{Q}} \longrightarrow GL(V \otimes_{\overline{\mathbb{Q}},\iota} \overline{\mathbb{Q}}_l),$$

one for each pair (l, ι), where l is a prime and $\iota : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_l$, which satisfy the following conditions.

- There is a multiset of integers $\mathrm{HT}(\mathcal{R})$ such that for each prime l and each embedding $\iota : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_l$ the restriction $R_{l,\iota}|_{G_{\mathbb{Q}_l}}$ is de Rham and $\mathrm{HT}(R_{l,\iota}|_{G_{\mathbb{Q}_l}}) = \mathrm{HT}(\mathcal{R})$.
- There is a finite set of primes S such that if $p \notin S$ then $\mathrm{WD}_p(R_{l,\iota})$ is unramified for all l and ι.
- For all but finitely many primes p there is a Frobenius semi-simple WD-representation $\mathrm{WD}_p(\mathcal{R})$ over $\overline{\mathbb{Q}}$ such that for all primes $l \neq p$ and for all ι we have
$$\mathrm{WD}_p(R_{l,\iota})^{\mathrm{ss}} \sim \mathrm{WD}_p(\mathcal{R}).$$

We make the following subsidiary definitions.

- We call $\mathcal{R}$ *strongly compatible* if the last condition (the existence of $\mathrm{WD}_p(\mathcal{R})$) holds for all primes p.
- We call $\mathcal{R}$ *irreducible* if each $R_{l,\iota}$ is irreducible.
- We call $\mathcal{R}$ *pure* of weight $w \in \mathbb{Z}$, if for all but finitely many p and for all eigenvalues α of $r_p(\mathrm{Frob}_p)$, where $\mathrm{WD}_p(\mathcal{R}) = (r_p, N_p)$, we have
$$|\iota\alpha|_\infty^2 = p^w$$
for all embeddings $\iota : \overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$.
- We call $\mathcal{R}$ *geometric* if there is a smooth projective variety $X/\mathbb{Q}$ and integers $i \geq 0$ and j and a subspace
$$W \subset H^i(X(\mathbb{C}), \overline{\mathbb{Q}})$$
such that for all l and ι, $W \otimes_{\overline{\mathbb{Q}},\iota} \overline{\mathbb{Q}}_l$ is $G_{\mathbb{Q}}$ invariant and realises $R_{l,\iota}$.

Conjectures 1.1 and 1.2 lead one to make the following conjecture.

Conjecture 1.3 *1. If $R : G_{\mathbb{Q}} \to GL_n(\overline{\mathbb{Q}}_l)$ is a continuous semi-simple de Rham representation unramified at all but finitely many primes then R is part of a weakly compatible system.*

2. Any weakly compatible system is strongly compatible.

3. Any irreducible weakly compatible system $\mathcal{R}$ is geometric and pure of weight $(2/\dim \mathcal{R}) \sum_{h \in \mathrm{HT}(\mathcal{R})} h$.

A famous theorem of Cebotarev asserts that if $K/\mathbb{Q}$ is any Galois extension in which all but finitely many primes are unramified (i.e. for all but finitely many primes p the image of $I_{\mathbb{Q}_p}$ in $\mathrm{Gal}(K/\mathbb{Q})$ is trivial) then the Frobenius elements at unramified primes $\mathrm{Frob}_p \in \mathrm{Gal}(K/\mathbb{Q})$ are dense in $\mathrm{Gal}(K/\mathbb{Q})$. It follows that an

irreducible weakly compatible system $\mathcal{R}$ is uniquely determined by $\mathrm{WD}_p(\mathcal{R})$ for all but finitely many p and hence by one $R_{l,\iota}$.

Conjectures 1.1 and 1.3 are known for one dimensional representations, in which case they have purely algebraic proofs based on class field theory (see [Se]). Otherwise only fragmentary cases have been proved, where amazingly the arguments are extremely indirect involving sophisticated analysis and geometry. We will come back to this later.

2. *L*-functions

L-functions are certain Dirichlet series

$$\sum_{n=1}^{\infty} a_n/n^s$$

which play an important role in number theory. A full discussion of the role of L-functions in number theory is beyond the scope of this talk. The simplest example of an L-function is the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} 1/n^s.$$

It converges to a holomorphic function in the half plane $\mathrm{Re}\, s > 1$ and in this region of convergence it can also be expressed as a convergent infinite product over the prime numbers

$$\zeta(s) = \prod_p (1 - 1/p^s)^{-1}.$$

This is called an *Euler product* and the individual factors are called Euler factors. Lying deeper is the fact that $\zeta(s)$ has meromorphic continuation to the whole complex plane, with only one pole: a simple pole at $s = 1$. Moreover if we set

$$Z(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$$

then Z satisfies the functional equation

$$Z(1 - s) = Z(s).$$

Encoded in the Riemann zeta function is lots of deep arithmetic information. For instance the location of the zeros of $\zeta(s)$ is intimately connected with the distribution of prime numbers. Moreover its special values at negative integers (where it is only defined by analytic continuation) turn out to be rational numbers encoding deep arithmetic information about the cyclotomic fields $\mathbb{Q}(e^{2\pi\sqrt{-1}/p})$.

Another celebrated example is the L-function of an elliptic curve E:

$$y^2 = x^3 + ax + b.$$

In this case the L-function is defined as an Euler product (converging in $\mathrm{Re}\, s > 3/2$)

$$L(E,s) = \prod_p L_p(E, p^{-s}),$$

where $L_p(E,X)$ is a rational function, and for all but finitely many p

$$L_p(E,X) = (1 - a_p(E)X + pX^2)^{-1},$$

with $p - a_p(E)$ being the number of solutions to the congruence

$$y^2 \equiv x^3 + ax + b \bmod p$$

in $\mathbb{F}_p^2$. It has recently been proved [BCDT] that $L(E,s)$ can be continued to an entire function, which satisfies a functional equation

$$(2\pi)^{-s}\Gamma(s)L(E,s) = \pm N(E)^{1-s}(2\pi)^{s-2}\Gamma(2-s)L(E,2-s),$$

for some explicit positive integer $N(E)$. A remarkable conjecture of Birch and Swinnerton-Dyer [BSD] predicts that $y^2 = x^3 + ax + b$ has infinitely many rational solutions if and only if $L(E,1) = 0$. Again we point out that it is the behaviour of the L-function at a point where it is only defined by analytic continuation, which is governing the arithmetic of E. This conjecture has been proved (see [Kol]) when $L(E,s)$ has at most a simple zero at $s = 1$.

One general setting in which one can define L-functions is l-adic representations. Let us look first at the local setting. If (r,N) is a WD-representation of $W_{\mathbb{Q}_p}$ on an E-vector space V, where E is an algebraically closed field of characteristic zero, we define a local L-factor

$$L((r,N),X) = \det(1 - X\mathrm{Frob}_p)|^{-1}_{V^{I_{\mathbb{Q}_p},N=0}} \in E(X).$$

($V^{I_{\mathbb{Q}_p},N=0}$ is the subspace of V where $I_{\mathbb{Q}_p}$ acts trivially and $N = 0$.) One can also associate to (r,N) a conductor $f(r,N) \in \mathbb{Z}_{\geq 0}$, which measures how deeply into $I_{\mathbb{Q}_p}$ the WD-representation (r,N) is nontrivial, and a local epsilon factor $\epsilon((r,N),\Psi_p) \in E$, which also depends on the choice of a non-trivial character $\Psi_p : \mathbb{Q}_p \to E^\times$ with open kernel. (See [Tat].)

If $R : G_{\mathbb{Q}} \to GL(V)$ is an l-adic representation of $G_{\mathbb{Q}}$ which is de Rham at l and pure of some weight $w \in \mathbb{Z}$, and if $\iota : \overline{\mathbb{Q}}_l \hookrightarrow \mathbb{C}$ we will define an L-function

$$L(\iota R, s) = \prod_p L(\iota \mathrm{WD}_p(R), p^{-s}),$$

which will converge to a holomorphic function in $\mathrm{Re}\, s > 1 + w/2$. For example

$$L(1,s) = \zeta(s)$$

and if $E/\mathbb{Q}$ is an elliptic curve then

$$L(\iota H^1(E(\mathbb{C}), \overline{\mathbb{Q}}_l), s) = L(E,s)$$

(for any ι). Note the useful formulae

$$L(\iota(R_1 \oplus R_2), s) = L(\iota R_1, s)L(\iota R_2, s) \qquad \text{and} \qquad L(\iota(R \otimes \chi_l^r), s) = L(\iota R, s + r).$$

Also note that $L(\iota R, s)$ determines $L(\mathrm{WD}_p(R), X)$ for all p and hence $\mathrm{WD}_p(R)^{\mathrm{ss}}$ for all but finitely many p. Hence by the Cebotarev density theorem $L(\iota R, s)$ determines R (up to semisimplification).

Write m_i^R for the multiplicity of an integer i in $\mathrm{HT}(R)$ and, if $w/2 \in \mathbb{Z}$, define $m_{w/2,\pm}^R \in (1/2)\mathbb{Z}$ by:

$$\begin{aligned} m_{w/2,+}^R + m_{w/2,-}^R &= m_{w/2}^R \\ m_{w/2,+}^R - m_{w/2,-}^R &= (-1)^{w/2}(\dim V^{c=1} - \dim V^{c=-1}). \end{aligned}$$

Assume that $m_{w/2,\pm}^R$ are integers, i.e. that $m_{w/2}^R \equiv \dim V \bmod 2$. Then we can define a Γ-factor, $\Gamma(R, s)$, which is a product of functions $\pi^{-(s+a)/2}\Gamma((s+a)/2)$ as a runs over a set of integers depending only on the numbers m_i^R and $m_{w/2,\pm}^R$. We can also define an epsilon factor $\epsilon_\infty(R, \Psi_\infty) \in \mathbb{C}^\times$ which again only depends on m_i^R, $m_{w/2,\pm}^R$ and a non-trivial character $\Psi_\infty : \mathbb{R} \to \mathbb{C}^\times$. Set

$$\Lambda(\iota R, s) = \Gamma(R, s)L(\iota R, s)$$

and

$$N(R) = \prod_p p^{f(\mathrm{WD}_p(R))}$$

(which makes sense as $f(\mathrm{WD}_p(R)) = 0$ for all but finitely many p) and

$$\epsilon(\iota R) = \epsilon_\infty(R, e^{2\pi\sqrt{-1}x}) \prod_p \iota\epsilon(\mathrm{WD}_p(R), \Psi_p),$$

where $\iota\Psi_p(x) = e^{-2\pi\sqrt{-1}x}$.

The following conjecture is a combination of conjecture 1.1 and conjectures which have become standard.

Conjecture 2.1 *Suppose that R is an irreducible l-adic representation of $G_{\mathbb{Q}}$ which is de Rham and pure of weight $w \in \mathbb{Z}$. Then $m_p^R = m_{w-p}^R$ for all p, so that $m_{w/2} \equiv \dim V \bmod 2$. Moreover the following should hold.*

1. *$L(\iota R, s)$ extends to an entire function, except for a single simple pole if $R = \chi_l^{-w/2}$.*
2. *$\Lambda(\iota R, s)$ is bounded in vertical strips $\sigma_0 \leq \mathrm{Re}\, s \leq \sigma_1$.*
3. *$\Lambda(\iota R, s) = \epsilon(\iota R)N(R)^{-s}\Lambda(\iota R^\vee, 1 - s)$.*

It is tempting to believe that something like properties 1., 2. and 3. should characterise those Euler products which arise from l-adic representations. We will discuss a more precise conjecture along these lines in the next section. Why Galois representations should be **the** source of Euler products with good functional equations seems a complete mystery.

3. Automorphic forms

Automorphic forms may be thought of as certain smooth functions on the quotient $GL_n(\mathbb{Z})\backslash GL_n(\mathbb{R})$. We need several preliminaries before we can make a precise definition.

Let $\widehat{\mathbb{Z}}$ denote the profinite completion of $\mathbb{Z}$, i.e.

$$\widehat{\mathbb{Z}} = \lim_{\leftarrow N} \mathbb{Z}/N\mathbb{Z} = \prod_p \mathbb{Z}_p,$$

a topological ring. Also let $\mathbb{A}^\infty$ denote the topological ring of finite adeles

$$\mathbb{A}^\infty = \widehat{\mathbb{Z}} \otimes_{\mathbb{Z}} \mathbb{Q},$$

where $\widehat{\mathbb{Z}}$ is an open subring with its usual topology. As an abstract ring, $\mathbb{A}^\infty$ is the subring of $\prod_p \mathbb{Q}_p$ consisting of elements (x_p) with $x_p \in \mathbb{Z}_p$ for all but finitely many p. However the topology is not the subspace topology. We define the topological ring of adeles to be the product

$$\mathbb{A} = \mathbb{A}^\infty \times \mathbb{R}.$$

Note that $\mathbb{Q}$ embeds diagonally as a discrete subring of $\mathbb{A}$ with compact quotient

$$\mathbb{Q}\backslash\mathbb{A} = \widehat{\mathbb{Z}} \times \mathbb{Z}\backslash\mathbb{R}.$$

We will be interested in $GL_n(\mathbb{A})$, the locally compact topological group of $n \times n$ invertible matrices with coefficients in $\mathbb{A}$. We remark that the topology on $GL_n(\mathbb{A})$ is the subspace topology resulting from the closed embedding

$$\begin{array}{ccc} GL_n(\mathbb{A}) & \hookrightarrow & M_n(\mathbb{A}) \times M_n(\mathbb{A}) \\ g & \mapsto & (g, g^{-1}). \end{array}$$

$GL_n(\mathbb{Q})$ is a discrete subgroup of $GL_n(\mathbb{A})$ and the quotient $GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A})$ has finite volume. If $U \subset GL_n(\widehat{\mathbb{Z}})$ is an open subgroup with $\det U = \widehat{\mathbb{Z}}^\times$, then

$$GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A})/U = (GL_n(\mathbb{Q}) \cap U)\backslash GL_n(\mathbb{R}).$$

Note that $GL_n(\mathbb{Q}) \cap U$ is a subgroup of $GL_n(\mathbb{Z})$ of finite index. Most of the statements we make concerning $GL_n(\mathbb{A})$ can be rephrased to involve only $GL_n(\mathbb{R})$, but at the expense of making them much more cumbersome. To achieve brevity (and because it seems more natural) we have opted to use the language of adeles. We hope that this extra abstraction will not be too confusing for the novice.

Before continuing our introduction of automorphic forms let us digress to mention class field theory, which provides a concrete example of the presentational advantages of the adelic language. It also implies essentially all the conjectures we are considering in the case of one dimensional Galois representations. Indeed this article is about the search for a non-abelian analogue of class field theory. Class field theory gives a concrete description of the abelianisation (maximal continuous

abelian quotient) $G_{\mathbb{Q}}^{\mathrm{ab}}$ of $G_{\mathbb{Q}}$ and $W_{\mathbb{Q}_p}^{\mathrm{ab}}$ of $W_{\mathbb{Q}_p}$. First the local theory asserts that there is an isomorphism

$$\mathrm{Art}_p : \mathbb{Q}_p^\times \xrightarrow{\sim} W_{\mathbb{Q}_p}^{\mathrm{ab}}$$

with various natural properties, including the facts that $\mathrm{Art}(\mathbb{Z}_p^\times)$ is the image of the inertia group $I_{\mathbb{Q}_p}$ in $W_{\mathbb{Q}_p}^{\mathrm{ab}}$, and that the induced map

$$\mathbb{Q}_p^\times/\mathbb{Z}_p^\times \longrightarrow W_{\mathbb{Q}_p}^{\mathrm{ab}}/I_{\mathbb{Q}_p} \subset G_{\mathbb{F}_p}$$

takes p to the geometric Frobenius element Frob_p. Secondly the global theory asserts that there is an isomorphism

$$\mathrm{Art} : \mathbb{A}^\times/\mathbb{Q}^\times \mathbb{R}_{>0}^\times \xrightarrow{\sim} G_{\mathbb{Q}}^{\mathrm{ab}}$$

such that the restriction of Art to $\mathbb{Q}_p^\times$ coincides with the composition of Art_p with the natural map $W_{\mathbb{Q}_p}^{\mathrm{ab}} \to G_{\mathbb{Q}}^{\mathrm{ab}}$. Thus Art is defined completely from a knowledge of the Art_p (and the fact that Art takes $-1 \in \mathbb{R}^\times$ to complex conjugation) and the reciprocity theorem of global class field theory can be thought of as a determination of the kernel of $\prod_p \mathrm{Art}_p$.

We now return to our (extended) definition of automorphic forms. For each partition $n = n_1 + n_2$ let N_{n_1,n_2} denote the subgroup of GL_n consisting of matrices of the form

$$\begin{pmatrix} I_{n_1} & * \\ 0 & I_{n_2} \end{pmatrix}.$$

Let $O(n) \subset GL_n(\mathbb{R})$ denote the orthogonal subgroup. Let $\mathfrak{z}_n$ denote the centre of the universal enveloping of $\mathfrak{gl}_n$, the complexified Lie algebra of $GL_n(\mathbb{R})$ (i.e. $\mathfrak{gl}_n = M_n(\mathbb{C})$ with $[X,Y] = XY - YX$). Via the Harish-Chandra isomorphism (see for example [Dix]) we may identify homomorphisms $\mathfrak{z}_n \to \mathbb{C}$ with multisets of n complex numbers. We will write χ_H for the homomorphism corresponding to a multiset H. Thus $\mathfrak{z}_n$ acts on the irreducible finite dimensional $\mathfrak{gl}_n$-module with highest weight $(a_1, ..., a_n) \in \mathbb{Z}^n$ $(a_1 \geq ... \geq a_n)$ by $\chi_{\{a_1+(n-1)/2,...,a_n+(1-n)/2\}}$.

Fix such a multiset H of cardinality n. The space of cusp forms with infinitesimal character H, $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$ is the space of smooth bounded functions

$$f : GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}) \longrightarrow \mathbb{C}$$

satisfying the following conditions.

1. (K-finiteness) The translates of f under $GL_n(\widehat{\mathbb{Z}}) \times O(n)$ (where $O(n)$ denotes the orthogonal group) span a finite dimensional vector space;
2. (Infinitesimal character H) If $z \in \mathfrak{z}_n$ then $zf = \chi_H(z)f$;
3. (Cuspidality) For each partition $n = n_1 + n_2$,

$$\int_{N_{n_1,n_2}(\mathbb{Q})\backslash N_{n_1,n_2}(\mathbb{A})} f(ug)du = 0.$$

Note that if $U \subset GL_n(\widehat{\mathbb{Z}})$ is an open subgroup with $\det U = \widehat{\mathbb{Z}}^\times$ then one may think of $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))^U$ as a space of functions on $(GL_n(\mathbb{Q}) \cap U)\backslash GL_n(\mathbb{R})$.

One would like to study $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$ as a representation of $GL_n(\mathbb{A})$, unfortunately it is not preserved by the action of $GL_n(\mathbb{R})$ (because the K-finiteness condition depends on the choice of a maximal compact subgroup $O(n) \subset GL_n(\mathbb{R})$). It does however have an action of $GL_n(\mathbb{A}^\infty) \times O(n)$ and of $\mathfrak{gl}_n$, which is essentially as good. More precisely it is an admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-module in the sense of [Fl]. In fact it is a direct sum of irreducible, admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-modules each occurring with multiplicity one. We will (slightly abusively) refer to these irreducible constituents as cuspidal automorphic representations of $GL_n(\mathbb{A})$ with infinitesimal character H.

$\mathcal{A}_{\{0\}}^\circ(\mathbb{Q}^\times \backslash \mathbb{A}^\times)$ is just the space of locally constant functions on $\mathbb{A}^\times/\mathbb{Q}^\times \mathbb{R}_{>0}^\times$ and so cuspidal automorphic representations of $GL_1(\mathbb{A})$ with infinitesimal character $\{0\}$, are just the (finite order) complex valued characters of $\mathbb{A}^\times/\mathbb{Q}^\times \mathbb{R}_{>0}^\times \cong \widehat{\mathbb{Z}}^\times$, i.e. Dirichlet characters. $\mathcal{A}_{\{s\}}^\circ(\mathbb{Q}^\times \backslash \mathbb{A}^\times)$ is simply obtained from $\mathcal{A}_{\{0\}}^\circ(\mathbb{Q}^\times \backslash \mathbb{A}^\times)$ by twisting by $|\ |^s$, where $|\ | : \mathbb{A}^\times/\mathbb{Q}^\times \to \mathbb{R}_{>0}^\times$ is the product of the absolute values $|\ |_x$. Thus in the case $n = 1$ cuspidal automorphic representations are essentially Dirichlet characters.

The case $n = 2$ is somewhat more representative. In this case we have $\mathcal{A}_{\{s,t\}}^\circ(GL_2(\mathbb{Q})\backslash GL_2(\mathbb{A})) = (0)$ unless $s - t \in i\mathbb{R}$, $s - t \in \mathbb{Z}$ or $s - t \in (-1, 1)$. It is conjectured that the third possibility can not arise unless $s = t$. Let us consider the case $s - t \in \mathbb{Z}_{>0}$ a little further. If $s - t \in \mathbb{Z}_{>0}$ then it turns out that the irreducible constituents of $\mathcal{A}_{\{s,t\}}^\circ(GL_2(\mathbb{Q})\backslash GL_2(\mathbb{A}))$ are in bijection with the weight $1 + s - t$ holomorphic cusp forms on the upper half plane, which are normalised newforms (see for example [Mi]). Thus in some sense cuspidal automorphic representations are are also generalisations of classical holomorphic normalised newforms.

Note that if ψ is a character of $\mathbb{A}^\times/\mathbb{Q}^\times \mathbb{R}_{>0}^\times$ and if π is a cuspidal automorphic representation of $GL_n(\mathbb{A})$ with infinitesimal character H then $\pi \otimes (\psi \circ \det)$ is also a cuspidal automorphic representation with infinitesimal character H and the contragredient (dual) π^* of π is a cuspidal automorphic representation with infinitesimal character $-H = \{-h : h \in H\}$.

One of the main questions in the theory of automorphic forms is to describe the irreducible constituents of $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$. If we are to do this we first need some description of all irreducible admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-modules, and then we can try to say which occur in $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$.

Just as a character $\psi : \mathbb{A}^\times \to \mathbb{C}^\times$ can be factored as

$$\psi = \psi_\infty \times \prod_p \psi_p$$

where $\psi_p : \mathbb{Q}_p^\times \to \mathbb{C}^\times$ (resp. $\psi_\infty : \mathbb{R}^\times \to \mathbb{C}^\times$), so an irreducible, admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-module can be factorised as a restricted tensor product (see [Fl])

$$\pi \cong \bigotimes_x{}' \pi_x,$$

where π_∞ is an irreducible, admissible $(\mathfrak{gl}_n, O(n))$-module (see for example [Wa]), and each π_p is an irreducible smooth (i.e. stabilisers of vectors are open in $GL_n(\mathbb{Q}_p)$) representation of $GL_n(\mathbb{Q}_p)$ with $\pi_p^{GL_n(\mathbb{Z}_p)} \neq (0)$ for all but finitely many p. To the factors π_x one can associate various invariants (see [J]).

- A central character $\psi_x : \mathbb{Q}_x^\times \to \mathbb{C}^\times$.
- L-factors $L(\pi_p, X) \in \mathbb{C}(X)$.
- A Γ-factor $\Gamma(\pi_\infty, s)$.
- Conductors $f(\pi_p) \in \mathbb{Z}_{\geq 0}$.
- For each non-trivial character $\Psi_x : \mathbb{Q}_x \to \mathbb{C}^\times$ an epsilon factor $\epsilon(\pi_x, \Psi_x) \in \mathbb{C}^\times$.

(We also remark that $\mathfrak{z}_n$ acts via a character $\mathfrak{z}_n \to \mathbb{C}$ on any irreducible, admissible $(\mathfrak{gl}_n, O(n))$-module π_∞. This character is called the infinitesimal character of π_∞.) Now we may attach to π

- a central character $\psi_\pi = \prod_x \psi_x : \mathbb{A}^\times \to \mathbb{C}^\times$;
- an L-function $L(\pi, s) = \prod_p L(\pi_p, p^{-s})$ (which may or may not converge);
- an extended L-function $\Lambda(\pi, s) = \Gamma(\pi_\infty, s) L(\pi, s)$;
- a conductor $N(\pi) = \prod_p p^{f(\pi_p)}$ (which makes sense because $f(\pi_p) = 0$ when $\pi_p^{GL_n(\mathbb{Z}_p)} \neq (0)$);
- and an epsilon constant $\epsilon(\pi) = \prod_x \epsilon(\pi_x, \Psi_x) \in \mathbb{C}^\times$ where $\prod_x \Psi_x : \mathbb{A}/\mathbb{Q} \to \mathbb{C}^\times$ is any non-trivial character.

The following theorem and conjecture describe the (expected) relationship between automorphic forms and L-functions with Euler product and functional equation. We suppose $n > 1$. A similar theorem to theorem 3.1 is true for $n = 1$, except that $L(\pi, s)$ may have one simple pole. In this case it was due to Dirichlet. Conjecture 3.2 becomes vacuous if $n = 1$.

Theorem 3.1 (Godement-Jacquet, [GJ]) *Suppose that π is an irreducible constituent of $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$ with $n > 1$. Then $L(\pi, s)$ converges to a holomorphic function in some right half complex plane $\operatorname{Re} s > \sigma$ and can be continued to a holomorphic function on the whole complex plane so that $\Lambda(\pi, s)$ is bounded in all vertical strips $\sigma_1 \geq \operatorname{Re} s \geq \sigma_2$. Moreover $\Lambda(\pi, s)$ satisfies the functional equation*

$$\Lambda(\pi, s) = \epsilon(\pi) N(\pi)^{-s} \Lambda(\pi^*, 1 - s).$$

Conjecture 3.2 (Cogdell-Piatetski-Shapiro, [CPS1]) *Suppose that π is an irreducible, admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-module such that the central character of π is trivial on $\mathbb{Q}^\times$ and such that $L(\pi, s)$ converges in some half plane. Suppose also that for all characters $\psi : \mathbb{A}^\times / \mathbb{Q}^\times \mathbb{R}_{>0}^\times \to \mathbb{C}^\times$ the L-function $\Lambda(\pi \otimes (\psi \circ \det), s)$ (which will then converge in some right half plane) can be continued to a holomorphic function on the entire complex plane, which is bounded in vertical strips and satisfies a functional equation*

$$\Lambda(\pi \otimes (\psi \circ \det), s) = \epsilon(\pi \otimes (\psi \circ \det)) N(\pi \otimes (\psi \circ \det))^{-s} \Lambda(\pi^* \otimes (\psi^{-1} \circ \det), 1 - s).$$

($\Lambda(\pi^* \otimes (\psi^{-1} \circ \det), s)$ also automatically converges in some right half plane.) Then there is a partition $n = n_1 + \ldots + n_r$ and cuspidal automorphic representations π_i

of $GL_{n_i}(\mathbb{A})$ *such that*

$$\Lambda(\pi, s) = \prod_{i=1}^{r} \Lambda(\pi_i, s).$$

This conjecture is known to be true for $n = 2$ ([We], [JL]) and $n = 3$ ([JPSS]). For $n > 3$ a weaker form of this conjecture involving twisting by higher dimensional automorphic representations is known to hold (see [CPS1], [CPS2]). These results are called 'converse theorems'.

The reason for us introducing automorphic forms is because of a putative connection to Galois representations, which we will now discuss. But first let us discuss the local situation. It has recently been established ([HT], [He], [Ha]) that there is a natural bijection, rec_p, from irreducible smooth representations of $GL_n(\mathbb{Q}_p)$ to n-dimensional Frobenius semi-simple WD-representations of $W_{\mathbb{Q}_p}$ over $\mathbb{C}$. The key point here is that the bijection should be natural. We will not describe here exactly what this means (instead we refer the reader to the introduction to [HT]). It does satisfy the following.

- $\psi_\pi \circ \mathrm{Art}_p^{-1} = \det \mathrm{rec}_p(\pi)$, where ψ_π is the central character of π.
- $L(\mathrm{rec}_p(\pi), X) = L(\pi, X)$.
- $f(\mathrm{rec}_p(\pi)) = f(\pi)$.
- $\epsilon(\mathrm{rec}_p(\pi), \Psi_p) = \epsilon(\pi, \Psi_p)$ for any non-trivial character $\Psi_p : \mathbb{Q}_p \to \mathbb{C}^\times$.

The existence of rec_p can be seen as a non-abelian generalisation of local class field theory, as in the case $n = 1$ we have $\mathrm{rec}_p(\pi) = \pi \circ \mathrm{Art}_p^{-1}$.

Now suppose that $\iota : \overline{\mathbb{Q}}_l \to \mathbb{C}$ and that R is a de Rham l-adic representation of $G_{\mathbb{Q}}$ which is unramified at all but finitely many primes. Using the local reciprocity map rec_p, we can associate to R an irreducible, admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-module

$$\pi(\iota R) = \pi_\infty(R) \otimes \prod_p \mathrm{rec}_p^{-1}(\iota \mathrm{WD}_p(R)),$$

where $\pi_\infty(R)$ is a tempered irreducible, admissible $(\mathfrak{gl}_n, O(n))$-module with infinitesimal character $\mathrm{HT}(R|_{G_{\mathbb{Q}_l}})$ and with $\Gamma(\pi_\infty(R), s) = \Gamma(R, s)$. The definition of $\pi_\infty(R)$ depends only on the numbers m_i^R and $m_{w/2,\pm}^R$. Then we have the following conjectures.

Conjecture 3.3 *Suppose that H is a multiset of n* integers *and that π is an irreducible constituent of $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$. Identify $\overline{\mathbb{Q}} \subset \mathbb{C}$. Then each* $\mathrm{rec}_p(\pi_p)$ *can be defined over $\overline{\mathbb{Q}}$ and there is an irreducible geometric strongly compatible system of l-adic representations $\mathcal{R}$ such that* $\mathrm{HT}(\mathcal{R}) = H$ *and* $\mathrm{WD}_p(\mathcal{R})^{\mathrm{ss}} = \mathrm{rec}_p(\pi_p)$ *for all primes p.*

Conjecture 3.4 *Suppose that*

$$R : G_{\mathbb{Q}} \longrightarrow GL(V)$$

is an irreducible l-adic representation which is unramified at all but finitely many primes and for which $R|_{G_{\mathbb{Q}_l}}$ is de Rham. Let $\iota : \overline{\mathbb{Q}}_l \to \mathbb{C}$. Then $\pi(\iota R)$ is a cuspidal automorphic representation of $GL_n(\mathbb{A})$.

These conjectures are essentially due to Langlands [Lan1], except we have used a precise formulation which follows Clozel [Cl1] and we have incorporated conjecture 1.1 into conjecture 3.4.

Conjecture 3.4 is probably the more mysterious of the two, as only the case $n = 1$ and fragmentary cases where $n = 2$ are known. This will be discussed further in the next section. Note the similarity to the main theorem of global class field theory that $\prod_p \mathrm{Art}_p : \mathbb{A}^\times \to G_{\mathbb{Q}}^{\mathrm{ab}}$ has kernel $\mathbb{Q}^\times$.

The following theorem provides significant evidence for conjecture 3.3.

Theorem 3.5 ([Kot], [Cl2], [HT]) *Suppose that H is multiset of n* distinct *integers and that π is an irreducible constituent of $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$. Let $\iota : \overline{\mathbb{Q}}_l \hookrightarrow \mathbb{C}$. Suppose moreover that $\pi^* \cong \pi \otimes (\psi \circ \det)$ for some character $\psi : \mathbb{A}^\times/\mathbb{Q}^\times \to \mathbb{C}^\times$, and that either $n \leq 2$ or for some prime p the representation π_p is square integrable (i.e. $\mathrm{rec}_p(\pi_p)$ is indecomposable). Then there is a continuous representation*

$$R_{l,\iota} : G_{\mathbb{Q}} \longrightarrow GL_n(\overline{\mathbb{Q}}_l)$$

with the following properties.

1. *$R_{l,\iota}$ is geometric and pure of weight $2/n \sum_{h \in H} h$.*
2. *$R_{l,\iota}|_{G_{\mathbb{Q}_l}}$ is de Rham and $\mathrm{HT}(R_{l,\iota}|_{G_{\mathbb{Q}_l}}) = H$.*
3. *For any prime $p \neq l$ there is a representation $r_p : W_{\mathbb{Q}_p} \to GL_n(\overline{\mathbb{Q}}_l)$ such that $\mathrm{WD}_p(R_{l,\iota})^{\mathrm{ss}} = (r_p, N_p)$ and $\mathrm{rec}_p(\pi_p) = (\iota r_p, N_p')$.*

This was established by finding the desired l-adic representations in the cohomology of certain unitary group Shimura varieties. It seems not unreasonable to hope that similar techniques might allow one to improve many of the technical defects in the theorem. However Clozel has stressed that in the cases where H does not have distinct elements or where $\pi^* \ncong \pi \otimes (\psi \circ \det)$, there seems to be no prospect of finding the desired l-adic representations in the cohomology of Shimura varieties. It seems we need a new technique.

4. What do we know?

Let us first summarise in a slightly less precise way the various conjectures we have made, in order to bring together the discussion so far. Fix an embedding $\overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$ and let H be a multiset of integers of cardinality $n > 1$. Then the following sets should be in natural bijection. One way to make precise the meaning of 'natural' is to require that two objects M and M' should correspond if the local L-factors $L_p(M, X)$ and $L_p(M', X)$ are equal for all but finitely many p. Note that in each case the factors $L_p(M, X)$ for all but finitely many p, completely determine M.

(AF) Irreducible constituents π of $\mathcal{A}_H^\circ(GL_n(\mathbb{Q})\backslash GL_n(\mathbb{A}))$.

(LF) Near equivalence classes of irreducible, admissible $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-modules π with the following properties. (We call two $GL_n(\mathbb{A}^\infty) \times (\mathfrak{gl}_n, O(n))$-modules, π and π' nearly equivalent if $\pi_p \cong \pi_p'$ for all but finitely many p.)

(a) π_∞ has infinitesimal character H.
(b) The central character ψ_π is trivial on $\mathbb{Q}^\times \subset \mathbb{A}^\times$.
(c) For all characters $\psi : \mathbb{A}^\times/\mathbb{Q}^\times \mathbb{R}^\times_{>0}$ the L-function $\Lambda(\pi \otimes (\psi \circ \det), s)$ converges in some right half plane, has holomorphic continuation to the entire complex plane so that it is bounded in vertical strips and satisfies the functional equation

$$\Lambda(\pi \otimes \psi, s) = \epsilon(\pi \otimes \psi) N(\pi \otimes \psi)^{-s} \Lambda(\pi^* \otimes \psi^{-1}, 1-s).$$

(d) (See [JS] for an explanation of this condition.) There is a finite set of primes S containing all primes p for which $\mathrm{rec}_p(\pi_p)$ is ramified, such that, writing $L(\pi_p, X) = \prod_{i=1}^n (1 - \alpha_{p,i} X)^{-1}$ for $p \notin S$,

$$\sum_{p \notin S, i, j} \sum_{m=1}^{\infty} \alpha_{p,i}^m \alpha_{p,j}^{-m} / m p^{ms} + \log(s-1)$$

is bounded as $s \to 1$ from the right.

In this case $L_p(\pi, X) = L(\pi_p, X)$.

(lR) (Fix $\iota : \overline{\mathbb{Q}}_l \to \mathbb{C}$.) Irreducible l-adic representations

$$R : G_{\mathbb{Q}} \longrightarrow GL_n(\overline{\mathbb{Q}}_l)$$

which are unramified at all but finitely many primes and for which $R|_{G_{\mathbb{Q}_l}}$ is de Rham with $\mathrm{HT}(R|_{G_{\mathbb{Q}_l}}) = H$. In this case $L_p(R, X) = \iota L(\mathrm{WD}_p(R), X)$.

(WCS) Irreducible weakly compatible systems of l-adic representations $\mathcal{R}$ with
$\mathrm{HT}(\mathcal{R}) = H$. In this case $L_p(\mathcal{R}, X) = L_p(\mathrm{WD}_p(\mathcal{R}), X)$.

(GCS) Irreducible geometric strongly compatible systems of l-adic representations $\mathcal{R}$ with $\mathrm{HT}(\mathcal{R}) = H$. In this case $L_p(\mathcal{R}, X) = L_p(\mathrm{WD}_p(\mathcal{R}), X)$.

For $n = 1$ we must drop the item (LF), because it would need to be modified to allow $L(\pi \otimes (\psi \circ \det), s)$ to have a simple pole, while, in any case condition (LF) (b) would make the implication $(LF) \Longrightarrow (AF)$ trivial. This being said, in the case $n = 1$ all the other four sets are known to be in natural bijection (see [Se]). This basically follows because global class field theory provides an isomorphism

$$\mathrm{Art} : \mathbb{A}^\times/\mathbb{Q}^\times \mathbb{R}^\times_{>0} \xrightarrow{\sim} G_{\mathbb{Q}}^{\mathrm{ab}}.$$

I would again like to stress how different are these various sorts of objects and how surprising it is to me that there is any relation between them. Items (AF) and (LF) both concern representations of adele groups, but arising in rather different settings: either from the theory of discrete subgroups of Lie groups or from the

theory of L-functions with functional equation. Items (lR) and (WCS) arise from Galois theory and item (GCS) arises from geometry.

So what do we know about the various relationships for $n > 1$?

Not much. Trivially one has $(GCS) \Longrightarrow (WCS) \Longrightarrow (lR)$. The passage $(AF) \Longrightarrow (LF)$ is OK by theorem 3.1. As discussed in section 3 we have significant partial results in the directions $(LF) \Longrightarrow (AF)$ and $(AF) \Longrightarrow (GCS)$, but both seem to need new ideas. (Though I should stress that I am not really competent to discuss converse theorems.)

One way to establish the equivalence of all five items would be to complete the passages $(LF) \Longrightarrow (AF)$ and $(AF) \Longrightarrow (GCS)$ and to establish the passage $(lR) \Longrightarrow (AF)$. It is these implications which have received most study, though it should be pointed out that in the function field case the equivalence of the analogous objects was established by looking at the implications

$$(lR) \Longrightarrow (LF) \Longrightarrow (AF) \Longrightarrow (GCS).$$

(See [Laf]. It is the use of techniques from Grothendieck's l-adic cohomology to prove the first of these implications which is most special to function fields.) However it is striking that in the case of number fields all known implications from items (lR), (WCS) or (GCS) to (LF) go via (AF).

For the rest of this article we will concentrate on what still seems to be the least understood problem: the passage from (lR) or (WCS) to (AF) or (LF). Although the results we have are rather limited one should not underestimate their power. Perhaps the most striking illustration of this is that the lifting theorems discussed in section 4.4 (combined with earlier work using base change and converse theorems) allowed Wiles [Wi] to finally prove Fermat's last theorem.

The discussion in the rest of this paper will of necessity be somewhat more technical. In particular we will need to discuss automorphic forms, l-adic representations and so on over general number fields (i.e. fields finite over $\mathbb{Q}$). We will leave it to the reader's imagination exactly how such a generalisation is made. In this connection we should remark that if L/K is a finite extension of number fields and if R is a semi-simple de Rham l-adic representation of G_L which is unramified at all but finitely many primes, then (see [A])

$$L(R, s) = L(\mathrm{Ind}_{G_L}^{G_K} R, s)$$

(formally if the L-functions don't converge). In fact this is true Euler factor by Euler factor and similar results hold for conductors and ϵ-factors (see [Tat]). This observation can be extremely useful.

4.1 Cyclic base change

Suppose that G is a group, H a normal subgroup such that G/H is cyclic with generator of σ. It is an easy exercise that an irreducible representation r of H extends to a representation of G if and only if $r^\sigma \cong r$ as representations of H. If one believes conjectures 3.3 and 3.4, one might expect that if L/K is a cyclic Galois extension of number fields of prime order, if σ generates $\mathrm{Gal}\,(L/K)$ and if π is a

cuspidal automorphic representation of $GL_n(\mathbb{A}_L)$ with $\pi \circ \sigma \cong \pi$, then there should be a cuspidal automorphic representation Π of $GL_n(\mathbb{A}_K)$, such that for all places w of L we have $\mathrm{rec}_w(\pi_w) = \mathrm{rec}_{w|_K}(\Pi_{w|_K})|_{W_{L_w}}$. This is indeed the case. For $n = 1$ we have $\pi = \Pi \circ \mathbf{N}_{L/K}$, Langlands [Lan2] proved it for $n = 2$ using the trace formula and Arthur and Clozel [AC] generalised his method to all n.

One drawback of this result is that if v is a place of K inert in L then there is no complete recipe for Π_v in terms of π. This can be surprisingly serious. It can however be alleviated, if we know how to associate irreducible l-adic representations to both Π and π. Langlands used this to show that many two dimensional Artin representations (i.e. l-adic representations with finite image) were automorphic (i.e. associated to a cuspidal automorphic representation). In fact using additional results from the theory of L-functions, particularly the converse theorem for GL_3 (see section 4.3), he and Tunnell ([Tu]) were able to establish the automorphy of all continuous two dimensional Artin representations with soluble image.

4.2 Brauer's theorem

The result I want to discuss is a result of Brauer [Br] about finite groups.

Theorem 4.1 (Brauer) *Suppose that r is a representation of a finite group G. Then there are soluble subgroups $H_i < G$, one dimensional representations ψ_i of H_i and integers n_i such that as virtual representations of G we have*

$$r = \sum_i n_i \mathrm{Ind}_{H_i}^G \psi_i.$$

As Artin [A] had realised this theorem has the following immediate consequence. (Indeed Brauer proved his theorem in response to Artin's work.)

Corollary 4.2 *Let $\iota : \overline{\mathbb{Q}}_l \to \mathbb{C}$. Suppose that*

$$R : G_{\mathbb{Q}} \longrightarrow GL_n(\overline{\mathbb{Q}}_l)$$

is an l-adic representation with finite image. *Then the L-function $L(\iota R, s)$ has meromorphic continuation to the entire complex plane and satisfies the expected functional equation.*

Artin's argument runs as follows. Let G denote the image of R and write

$$R = \sum_i n_i \mathrm{Ind}_{H_i}^G \psi_i$$

as in Brauer's theorem. Let $L/\mathbb{Q}$ be the Galois extension with group G cut out by R and let $K_i = L^{H_i}$. Then one has equalities

$$\begin{aligned} L(\iota R, s) &= \textstyle\prod_i L(\iota \mathrm{Ind}_{G_{K_i}}^{G_{\mathbb{Q}}} \psi_i, s)^{n_i} \\ &= \textstyle\prod_i L(\iota \psi_i, s)^{n_i}. \end{aligned}$$

By class field theory for the fields K_i, the character ψ_i is automorphic on $GL_1(\mathbb{A}_{K_i})$ and so $L(\iota\psi_i, s)$ has holomorphic continuation to the entire complex plane (except possibly for one simple pole if $\psi_i = 1$) and satisfies a functional equation. It follows that $L(\iota R, s)$ has meromorphic continuation to the entire complex plane and satisfies a functional equation. The problem with this method is that some of the integers n_i will usually be negative so that one can only conclude the meromorphy of $L(\iota R, s)$, not its holomorphy.

4.3 Converse theorems

As Cogdell and Piatetski-Shapiro point out, conjecture 3.2 would have very important implications for Galois representations. For instance the cases $n = 2$ and 3 played a key role in the proof of the automorphy of two dimensional Artin representations (see 4.1). Conjecture 3.2 combined with Brauer's theorem and a result of Jacquet and Shalika [JS] in fact implies that many (all? - certainly those with soluble or perfect image) Artin representations are automorphic. A similar argument shows that in many other cases, in order to check the automorphy of an l-adic representation of $G_{\mathbb{Q}}$, it suffices to do so after a finite base change. For instance one has the following result.

Assume conjecture 3.2. Let $\iota : \overline{\mathbb{Q}}_l \hookrightarrow \mathbb{C}$ *and let* $K/\mathbb{Q}$ *be a finite, totally real Galois extension. Suppose that* Π *is a cuspidal automorphic representation of* $GL_n(\mathbb{A}_K)$ *with infinitesimal character corresponding to a multiset* H *consisting of* n *distinct integers. If* $n > 2$ *also suppose that* Π_v *is square integrable (i.e.* $\mathrm{rec}_v(\Pi_v)$ *is indecomposable) for some finite place* v *of* K*. Let*

$$R : G_{\mathbb{Q}} \longrightarrow GL_n(\overline{\mathbb{Q}}_l)$$

be an l*-adic representation such that* $R \sim R^* \otimes \psi$ *for some character* ψ *of* $G_{\mathbb{Q}}$*, and such that* $R|_{G_K}$ *is irreducible. Suppose finally that* $R|_{G_K}$ *and* Π *are associated, in the sense that, for all but finitely many places* v *of* K*, we have*

$$\iota L(\mathrm{WD}_v(R|_{G_K}), X) = L(\Pi_v, X).$$

Then there is a regular algebraic cuspidal automorphic representation π *of* $GL_n(\mathbb{A})$ *associated to* R *in the same sense.*

4.4 Lifting theorems

To describe this sort of theorem we first remark that if $R : G_{\mathbb{Q}} \to GL_n(\overline{\mathbb{Q}}_l)$ is continuous then after conjugating R by some element of $GL_n(\overline{\mathbb{Q}}_l)$ we may assume that the image of R is contained in $GL_n(\mathcal{O}_{\overline{\mathbb{Q}}_l})$ and so reducing we obtain a continuous representation

$$\overline{R} : G_{\mathbb{Q}} \longrightarrow GL_n(\overline{\mathbb{F}}_l).$$

The lifting theorems I have in mind are results of the general form if R and R' are l-adic representations of $G_{\mathbb{Q}}$ with R' automorphic and if $\overline{R} = \overline{R}'$ then R is also automorphic. Very roughly speaking the technique (pioneered by Wiles [Wi] and

completed by the author and Wiles [TW]) is to show that $R \bmod l^r$ arises from automorphic forms for all r by induction on r. As $\ker(GL_n(\mathbb{Z}/l^r\mathbb{Z}) \twoheadrightarrow GL_n(\mathbb{Z}/l^{r-1}\mathbb{Z}))$ is an abelian group one is led to questions of class field theory and Galois cohomology.

I should stress that such theorems are presently available only in very limited situations. I do not have the space to describe the exact limitations, which are rather technical, but the sort of restrictions that are common are as follows.

1. If $R : G_{\mathbb{Q}} \to GL(V)$ then there should be a character $\mu : G_{\mathbb{Q}} \to GL_n(\overline{\mathbb{Q}}_l)$ and a non-degenerate bilinear form $(\ ,\ )$ on V such that
   - $(R(\sigma)v_1, R(\sigma)v_2) = \mu(\sigma)(v_1, v_2)$ and
   - $(v_2, v_1) = \mu(c)(v_1, v_2)$.

   (This seems to be essential for the method of [TW].)
2. R should be de Rham with distinct Hodge-Tate numbers. (This again seems essential to the method of [TW], but see [BT].)
3. Either R and R' should be ordinary (i.e. their restrictions to $G_{\mathbb{Q}_l}$ should be contained in a Borel subgroup); or R and R' should be crytsalline (not just de Rham) at l with the same Hodge-Tate numbers and l should be large compared with the differences of elements of $\mathrm{HT}(R)$. (The problems here are connected with the need for an integral Fontaine theory, but they are not simply technical problems. There are some complicated results pushing back this restriction in isolated cases, see [CDT], [BCDT], [Sa], but so far our understanding is very limited.)
4. The image of $\overline{R}$ should not be too small (e.g. should be irreducible when restricted to $\mathbb{Q}(e^{2\pi i/l})$), though in the case $n = 2$ there is beautiful work of Skinner and Wiles ([SW1] and [SW3]) dispensing with this criterion, which this author has unfortunately not fully understood.

In addition, all the published work is for the case $n = 2$. However there is ongoing work of a number of people attempting to dispense with this assumption. Using a very important insight of Diamond [Dia], the author, together with L.Clozel and M.Harris, has generalised to all n the so called minimal case (originally treated in [TW]) where R is no more ramified than $\overline{R}$. One would hope to be able to deduce the non-minimal case from this, as Wiles did in [Wi] for $n = 2$. In this regard one should note the work of Skinner and Wiles [SW2] and the work of Mann [Ma]. However there seems to be one missing ingredient, the analogue of the ubiquitous Ihara lemma, see lemma 3.2 of [Ih] (and also theorem 4.1 of [R]). As this seems to be an important question, but one which lies in the theory of discrete subgroups of Lie groups, let us take the trouble to formulate it, in the hope that an expert may be able to prove it. It should be remarked that there are a number of possible formulations, which are not completely equivalent and any of which would seem to suffice. We choose to present one which has the virtue of being relatively simple to state.

Conjecture 4.3 *Suppose that $G/\mathbb{Q}$ is a unitary group which becomes an inner form of GL_n over an imaginary quadratic field E. Suppose that $G(\mathbb{R})$ is compact. Let l be a prime which one may assume is large compared to n. Let p_1 and p_2 be distinct primes different from l with $G(\mathbb{Q}_{p_1}) \cong GL_n(\mathbb{Q}_{p_1})$ and $G(\mathbb{Q}_{p_2}) \cong GL_n(\mathbb{Q}_{p_2})$.*

Let U be an open compact subgroup of $G(\mathbb{A}^{p_1,p_2})$ and consider the representation of $GL_n(\mathbb{Q}_{p_1}) \times GL_n(\mathbb{Q}_{p_2})$ on the space $C^\infty(G(\mathbb{Q})\backslash G(\mathbb{A})/U, \overline{\mathbb{F}}_l)$ of locally constant $\overline{\mathbb{F}}_l$-valued functions on

$$G(\mathbb{Q})\backslash G(\mathbb{A})/U = (G(\mathbb{Q}) \cap U)\backslash (GL_n(\mathbb{Q}_{p_1}) \times GL_n(\mathbb{Q}_{p_2}).$$

(Note that $G(\mathbb{Q}) \cap U$ is a discrete cocompact subgroup of $GL_n(\mathbb{Q}_{p_1}) \times GL_n(\mathbb{Q}_{p_2})$.) Suppose that $\pi_1 \otimes \pi_2$ is an irreducible sub-representation *of $C^\infty(G(\mathbb{Q})\backslash G(\mathbb{A})/U, \overline{\mathbb{F}}_l)$ with π_1 generic. Then π_2 is also generic.*

The most serious problem with applying such lifting theorems to prove an l-adic representation R is automorphic is the need to find some way to show that $\overline{R}$ is automorphic. The main success of lifting theorems to date, has been to show that if E is an elliptic curve over the rationals then $H^1(E(\mathbb{E}), \overline{\mathbb{Q}}_l)$ is automorphic, so that E is a factor of the Jacobian of a modular curve and the L-function $L(E,s)$ is an entire function satisfying the expected functional equation ([Wi], [TW],[BCDT]). This was possible because $GL_2(\mathbb{Z}_3)$ happens to be a pro-soluble group and there is a homomorphism $GL_2(\mathbb{F}_3) \longrightarrow GL_2(\mathbb{Z}_3)$ splitting the reduction map. The Artin representation

$$G_{\mathbb{Q}} \longrightarrow GL(H^1(E(\mathbb{C}), \mathbb{F}_3)) \longrightarrow GL_2(\mathbb{Z}_3)$$

is automorphic by the Langlands-Tunnell theorem alluded to in section 4.1.

4.5 Other techniques?

I would like to discuss one other technique which has been some help if $n = 2$ and may be helpful more generally. We will restrict our attention here to the case $n = 2$ and $\det R(c) = -1$. We have said that the principal problem with lifting theorems for proving an l-adic representation $R : G_{\mathbb{Q}} \to GL_2(\overline{\mathbb{Q}}_l)$ is automorphic is that one needs to know that $\overline{R}$ is automorphic. This seems to be a very hard problem. Nonetheless one can often show that $\overline{R}$ becomes automorphic over some Galois totally real field $K/\mathbb{Q}$. (Because K is totally real, if $\overline{R}(G_{\mathbb{Q}}) \supset SL_2(\mathbb{F}_l)$ and $l > 3$ then $\overline{R}(G_K) \supset SL_2(\mathbb{F}_l)$. So this 'potential automorphy' is far from vacuous). The way one does this is to look for an abelian variety A/K with multiplication by a number field F with $[F : \mathbb{Q}] = \dim A$, and such that $\overline{R}$ is realised on $H^1(A(\mathbb{C}), \mathbb{F}_l)[\lambda]$ for some prime $\lambda | l$, while for some prime $\lambda' | l' \neq l$ the image of G_K on $H^1(A(\mathbb{C}), \mathbb{F}_{l'})[\lambda']$ is soluble. One then argues that $H^1(A(\mathbb{C}), \mathbb{F}_{l'})[\lambda']$ is automorphic, hence by a lifting theorem $H^1(A(\mathbb{C}), \mathbb{Q}_{l'}) \otimes_{F_{l'}} F_{\lambda'}$ is automorphic, so that (tautologically) $H^1(A(\mathbb{C}), \mathbb{F}_l)[\lambda]$ is also automorphic, and hence, by another lifting theorem, $R|_{G_K}$ is automorphic. One needs K to be totally real, as over general number fields there seems to be no hope of proving lifting theorems, or even of attaching l-adic representations to automorphic forms. In practice, because of various limitations in the lifting theorems one uses, one needs to impose some conditions on the behaviour of a few primes, like l, in K and some other conditions on A. The problem of finding a suitable A over a totally real field K, comes down to finding a K-point on a twisted Hilbert modular variety. This is possible because we are free to choose K, the only restriction being that K is totally real and certain small

primes (almost) split completely in K. To do this, one has the following relatively easy result.

Proposition 4.4 ([MB],[P]) *Suppose that $X/\mathbb{Q}$ is a smooth geometrically irreducible variety. Let S be a finite set of places of $\mathbb{Q}$ and suppose that X has a point over the completion of $\mathbb{Q}$ at each place in S. Let $\mathbb{Q}_S$ be the maximal extension of $\mathbb{Q}$ in which all places in S split completely (e.g. $\mathbb{Q}_{\{\infty\}}$ is the maximal totally real field). Then X has a $\mathbb{Q}_S$-point.*

In this regard it would have extremely important consequences if, in the previous proposition, one could replace $\mathbb{Q}_S$ by $\mathbb{Q}_S^{\mathrm{sol}}$, the maximal soluble extension of $\mathbb{Q}$ in which all places in S split completely. I do not know if it is reasonable to expect this.

Using this method one can, for instance, prove the following result.

Theorem 4.5 ([Tay]) *Suppose that $\mathcal{R}$ is an irreducible weakly compatible system of two dimensional l-adic representations with $\mathrm{HT}(\mathcal{R}) = \{n_1, n_2\}$ where $n_1 \neq n_2$. Suppose also that $\det R_{l,\iota}(c) = -1$ for one (and hence for all) pairs (l, ι). Then there is a Galois totally real field $K/\mathbb{Q}$ and a cuspidal automorphic representation π of $GL_2(\mathbb{A}_K)$ such that*

- *for all $v|\infty$, π_v has infinitesimal character H, and*
- *for all (l, ι) and for all finite places $v \nmid l$ of K we have*

$$\mathrm{rec}_v(\pi_v) = \mathrm{WD}_v(R_{l,\iota}|_{G_K})^{\mathrm{ss}}.$$

In particular $\mathcal{R}$ is pure of weight $(n_1+n_2)/2$. Moreover $\mathcal{R}$ is strongly compatible and $L(\iota\mathcal{R}, s)$ has meromorphic continuation to the entire complex plane and satisfies the expected functional equation.

The last sentence of this theorem results from the first part and Brauer's theorem. We remark that conjecture 3.2 would imply that this theorem could be improved to assert the automorphy of $\mathcal{R}$ over $\mathbb{Q}$.

References

[A] E.Artin, *Zur Theorie der L-Reihen mit allgemeinen Gruppencharakteren*, Abh. Math. Sem. Univ. Hamburg 8 (1930), 292–306.

[AC] J.Arthur and L.Clozel, *Simple algebras, base change and the advanced theory of the trace formula*, Annals of Math. Studies 120, PUP 1989.

[Berg] L.Berger, *Représentations p-adiques et équations différentielles*, Invent. math. 148 (2002), 219–284.

[Bert] P.Berthelot, *Altérations de variétés algébriques (d'après A.J.de Jong)*, Astérisque 241 (1997), 273–311.

[BCDT] C.Breuil, B.Conrad, F.Diamond and R.Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$*, J.A.M.S. 14 (2001), 843–939.

[Br] R.Brauer, *On Artin's L-series with general group characters*, Ann. Math. (2) 48 (1947), 502–514.

[BSD] B.Birch and P.Swinnerton-Dyer, *Notes on elliptic curves II*, J. Reine Angew. Math. 218 (1965), 79–108.

[BT] K.Buzzard and R.Taylor, *Companion forms and weight one forms*, Annals of math 149 (1999), 905–919.

[CDT] B.Conrad, F.Diamond and R.Taylor, *Modularity of certain potentially Barsotti-Tate Galois representations*, JAMS 12 (1999), 521–567.

[Cl1] L.Clozel, *Motifs et formes automorphes: applications du principe de fonctorialité*, in "Automorphic forms, Shimura varieties and L-functions I", Academic Press 1990.

[Cl2] L.Clozel, *Représentations Galoisiennes associées aux representations automorphes autoduales de $GL(n)$*, Pub. Math. IHES 73 (1991), 97–145.

[CPS1] J.Cogdell and I.Piatetski-Shapiro, *Converse theorems for GL_n*, Publ. Math. IHES 79 (1994), 157–214.

[CPS2] J.Cogdell and I.Piatetski-Shapiro, *Converse theorems for GL_n II*, J. reine angew. Math. 507 (1999), 165–188.

[De] P.Deligne, *La conjecture de Weil I*, Publ. Math. IHES 43 (1974), 273–307.

[Dia] F.Diamond, *The Taylor-Wiles construction and multiplicity one*, Invent. Math. 128 (1997), 379–391.

[Dix] J.Dixmier, *Algèbres enveloppantes*, Gauthier Villars 1974.

[Fa] G.Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. math. 73 (1983), 349–366.

[Fl] D.Flath, *Decomposition of representations into tensor products*, in "Automorphic forms, representations and L-functions I" AMS 1979.

[Fo1] J.-M.Fontaine, talk at "Mathematische Arbeitstagung 1988", Maz-Planck-Institut für Mathematik preprint no. 30 of 1988.

[Fo2] J.-M.Fontaine, *Le corps des périodes p-adiques*, Astérisque 223 (1994), 59–111.

[Fo3] J.-M.Fontaine, *Représentations p-adiques semi-stables*, Astérisque 223 (1994), 113–184.

[Fo4] J.-M.Fontaine, *Représentations l-adiques potentiellement semi-stables*, Astérisque 223 (1994), 321–347.

[FM] J.-M.Fontaine and B.Mazur, *Geometric Galois representations*, in Elliptic curves, modular forms and Fermat's last theorem (J.Coates and S.-T.Yau eds.), International Press 1995.

[GJ] Godement and H.Jacquet, *Zeta functions of simple algebras*, LNM 260, Springer 1972.

[Ha] M.Harris, *On the local Langlands correspondence*, these proceedings.

[He] G.Henniart, *Une preuve simple des conjectures de Langlands pour $GL(n)$ sur un corps p-adiques*, Invent. Math. 139 (2000), 439–455.

[HT] M.Harris and R.Taylor, *The geometry and cohomology of some simple Shimura varieties*, PUP 2001.

[Ih] Y.Ihara, *On modular curves over finite fields*, in "Proceedings of an international colloquium on discrete subgroups of Lie groups and applications to moduli" OUP 1975.

[Il] L.Illusie, *Cohomologie de de Rham et cohomologie étale p-adique*,

Astérisque 189–190 (1990), 325–374.

[J] H.Jacquet, *Principal L-functions of the linear group*, in "Automorphic forms, representations and L-functions 2" AMS 1979.

[JL] H.Jacquet and R.Langlands, *Automorphic forms on $GL(2)$*, LNM 114, Springer 1970.

[JPSS] H.Jacquet, I.Piatetski-Shapiro and J.Shalika, *Automorphic forms on $GL(3)$*, Annals of math. 109 (1979), 169–258.

[JS] H.Jacquet and J.Shalika, *On Euler products and the classification of automorphic representations II*, Amer. J. Math. 103 (1981), 777–815.

[Kol] V.Kolyvagin, *On the Mordell-Weil group and the Shafarevich-Tate group of modular elliptic curves*, in "Proceedings of the Kyoto International Congress of Mathematicians" Math. Soc. Japan 1991.

[Kot] R.Kottwitz, *On the λ-adic representations associated to some simple Shimura varieties*, Invent. Math. 108 (1992), 653–665.

[Laf] L.Lafforgue, *Drinfeld varieties and the Langlands program*, these proceedings.

[Lan1] R.P.Langlands, *Automorphic representations, Shimura varieties and motives. Ein Märchen.* in "Automorphic forms, representations and L-functions II" AMS 1979.

[Lan2] R.P. Langlands, *Base change for $GL(2)$*, Annals of Math. Studies **96**, Princeton Univ. Press, Princeton, 1980.

[Ma] W.R.Mann, *Local level-raising for GL_n*, preprint.

[MB] L.Moret-Bailly, *Groupes de Picard et problèmes de Skolem II*, Ann. Sci. ENS 22 (1989), 181–194.

[Mi] T.Miyake, *Modular forms*, Springer 1989.

[P] F.Pop, *Embedding problems over large fields*, Annals of math 144 (1996), 1–34.

[R] K.Ribet, *Congruence relations between modular forms*, in "Proceedings of the Warsaw ICM" Polish Scientific Publishers 1984.

[Sa] D.Savitt, *modularity of some potentially Barsotti-Tate Galois representations*, preprint.

[Se] J.-P.Serre, *Abelian l-adic representations and elliptic curves*, Benjamin 1968.

[SW1] C.Skinner and A.Wiles, *Residually reducible representations and modular forms*, Inst. Hautes Etudes Sci. Publ. Math. 89 (1999), 5–126.

[SW2] C.Skinner and A.Wiles, *Base change and a problem of Serre*, Duke Math. J. 107 (2001), 15–25.

[SW3] C.Skinner and A.Wiles, *Nearly ordinary deformations of irreducible residual representations*, Annales de la Faculté de Sciences de Toulouse X (2001), 185–215.

[Tat] J.Tate, *Number theoretic background*, in A.Borel and W.Casselman "Automorphic forms, representations and L-functions", Proc. Symposia in Pure Math. 33 (2), AMS 1979.

[Tay] R.Taylor, *On the meromorphic continuation of degree two L-functions*, preprint available at `http://www.math.harvard.edu/~rtaylor`.

[Tu] J. Tunnell, *Artin's conjecture for representations of octahedral type*, Bull. AMS **5** (1981), 173–175.

[TW] R.Taylor and A.Wiles, *Ring theoretic properties of certain Hecke algebras*, Annals of math. 141 (1995), 553–572.

[Wa] N.Wallach, *Representations of reductive Lie groups*, in "Automorphic forms, representations and L-functions I" AMS 1979.

[We] A.Weil, *Über die Bestimmung Dirichletscher Reihen durch Funktionalgleichungen*, Math. Ann. 168 (1967), 149–156.

[Wi] A.Wiles, *Modular elliptic curves and Fermat's last theorem*, Annals of math. 141 (1995), 443–551.

ICM 2002 · Vol. I · 475–493

Geometry and Nonlinear Analysis*

Gang Tian†

Nonlinear analysis has played a prominent role in the recent developments in geometry and topology. The study of the Yang-Mills equation and its cousins gave rise to the Donaldson invariants and more recently, the Seiberg-Witten invariants. Those invariants have enabled us to prove a number of striking results for low dimensional manifolds, particularly, 4-manifolds. The theory of Gromov-Witten invariants was established by using solutions of the Cauchy-Riemann equation (cf. [RT], [LT], [FO], [Si], [Ru]). These solutions are often refered as pseudo-holomorphic maps which are special minimal surfaces studied long in geometry. It is certainly not the end of applications of nonlinear partial differential equations to geometry. In this talk, we will discuss some recent progress on nonlinear partial differential equations in geometry. We will be selective, partly because of my own interest and partly because of recent applications of nonlinear equations. There are also talks in this ICM to cover some other topics of geometric analysis by R. Bartnik, B. Andrew, P. Li and X.X. Chen, etc.

Standard partial differential equations in geometry are the Einstein equation, Yang-Mills equation, minimal surface equation as well as its close cousin, Harmonic map equation. There are also parabolic versions of these equations, leading to R. Hamilton's Ricci flow, the Yang-Mills flow and the mean curvature flow. The solutions, which played a fundamental role in geometry and topology, of these equations are their self-dual type ones. I will focus on self-dual type solutions in this talk. All these equations are in general hyperbolic equations if we allow Lorentz metrics on the underlying manifolds, but in differential geometry, so far, we only concern static solutions, that is, we assume that the metrics involved are Riemannian. I do believe that the study of this static case will be very important in our future understanding general Einstein equation.

1. Einstein equation

We will begin with the Einstein equation. We will always denote by M a differentiable manifold. A metric g on M is given by a non-degenerate matrix-

*Supported partially by NSF grants and a Simons fund.

†Department of Mathematics, Massachusetts Institute of Technology, USA and Beijing University, China. E-mail: tian@math.mit.edu

valued functions (g_{ij}) in local coordinates $x_1, \cdots, x_n$, where n is the dimension of M. Recall that g is Riemannian if the matrices (g_{ij}) are positive definite.

Associated to each metric, there is a canonical connection, the Levi-Civita connection, ∇ characterized by the torsion freeness and $\nabla g = 0$, which means that g is parallel. In local coordinates,

$$\nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_j} = \Gamma_{ij}^k \frac{\partial}{\partial x_k}, \quad \Gamma_{ij}^k = \frac{1}{2} g^{kl} \left(\frac{\partial g_{il}}{\partial x_j} + \frac{\partial g_{jl}}{\partial x_i} - \frac{\partial g_{ij}}{\partial x_l} \right), \tag{1.1}$$

where (g^{kl}) denotes the inverse of (g_{ij}). Then the curvature (R_{jkl}^i) is defined by

$$R_{jkl}^i = \frac{\partial \Gamma_{jk}^i}{\partial x_l} - \frac{\partial \Gamma_{jl}^i}{\partial x_k} + \Gamma_{sk}^i \Gamma_{jl}^s - \Gamma_{sl}^i \Gamma_{jk}^s. \tag{1.2}$$

The curvature is completely determined by its sectional curvatures, that is, Gauss curvatures of surface cross sections. The Ricci curvature $\mathrm{Ric} = (R_{ij})$ is given by taking trace of the curvature:

$$R_{ij} = \sum_k R_{ikj}^k. \tag{1.3}$$

The Ricci curvature essentially measures the variation of the volume form.

A metric g is called an Einstein metric if it satisfies the following Einstein equation

$$R_{ij} = \lambda g_{ij}, \tag{1.4}$$

where λ is a constant, usually called Einstein constant. (1.4) is the Euler-Lagrangian equation of the functional $\int_M s(g) dv$, where $s(g)$ denotes the scalar curvature of g, on the space of metrics with fixed volume. (1.4) is invariant under diffeomorphism group action. It is elliptic modulo diffeomorphisms when g is a Riemannian metric. From now on, I will assume that g is Riemannian.

The simplest examples of Einstein metrics include the euclidean metric on $\mathbb{R}^n$, the standard spherical metric on the unit sphere S^n and the hyperbolic metric on the unit ball $B^n \in \mathbb{R}^n$. In fact, all these metrics have constant sectional curvature 0 or 1 or -1, consequently, their Einstein constant λ are $0, n-1, -n+1$, respectively.

Every Riemannian 2-manifold (M, g) has a natural conformal structure. The classical Uniformization Theorem states that the universal covering of M together with the induced conformal structure is conformal to either $\mathbb{C}^1$ or S^2 or B^2 with canonical conformal structures. This implies that every Riemannian 2-manifold (M, g) admits a unique Einstein metric within its associated conformal class and with Einstein constant 0 or 1 or -1.

Another way of proving this existence is to use partial differential equation. Given a Riemannian 2-manifold (M, g), consider a metric $\tilde{g}$ conformal to g, so it is of the form $e^{\varphi} g$. A simple computation shows that $\tilde{g}$ is of constant curvature λ if and only if φ satisfies the following equation

$$\Delta \varphi + \frac{s(g)}{2} = \lambda e^{\varphi}. \tag{1.5}$$

This equation can be solved (cf. [Au], [KW]), so there is an Einstein metric in any given conformal class on any Riemannian 2-manifold. In early 1990's, R. Hamilton gave a heat flow proof of the Uniformization Theorem ([Ha], [Ch]). This new proof also yields a biproduct: The space of metrics on S^2 with positive curvature is contractible.

The uniformization for 2-manifolds led to two generalizations in higher dimensions. The one is the Yamabe problem (cf. [Au], [Sc]). The other is the Calabi's problem on Kähler-Einstein metrics which I will address more later.

For 3-manifolds, Einstein metrics are also of constant sectional curvature, so their universal coverings are either S^3 or $\mathbb{R}^3$ or hyperbolic 3-space H^3. A major part of Thurston's program is to show the existence of metrics with constant sectional curvature on 3-manifolds which satisfy certain mild topological conditions. Thurston claimed long time ago that an atoroidal Haken manifold admits a complete hyperbolic metric. It will be interesting to have an analytic proof of this claim by solving the corresponding Einstein equation. In general, one hopes that any 3-manifold can be canonically split into some pieces of simple topological type and other pieces which admit Einstein metrics. There are at least two possible approaches to this: one is the variational method, trying to minimax certain functional involving curvatures, while the other is to use the Ricci flow, hoping that one can understand how the singularity is formed along the flow. So far none of them work yet.

When the dimension is higher than or equal to 4, an Einstein metric may not be of constant sectional curvature. It is still a very interesting question to find out topological constraints on Einstein manifolds. If a 4-manifold M admits an Einstein metric, then the Hitchin-Thorpe inequality says that $|\tau(M)| \leq \frac{2}{3}|\chi(M)|$. This implies that the connected sum of more than 4 copies of $\mathbb{C}P^2$ can not have any Einstein metric. More recently, C. Lebrun showed that a 4-manifold M with non-vanishing Seiberg-Witten invariant admits an Einstein metric only if $3\tau(M) \leq \chi(M)$. We do not know any constraints on compact Einstein manifolds of dimension higher than 4. It may be possible that any manifold of dimension ≥ 5 has an Einstein metric.

Examples of Einstein metrics can be constructed by exploring symmetries, such as, homogeneous Einstein metrics, cohomogeneity one Einstein metrics. One can also construct new Einstein metrics from known ones through certain intrigue constructions when the underlying manifolds are of special fibration structures (cf. [Wang], [BG]).

When $n \geq 4$, there is a special class of solutions of the Einstein equation, that is, Einstein metrics of special holonomy. If (M, g) is an irreducible Riemannian manifold, a well-known theorem of M. Berger states that either (M, g) is a locally symmetric space or its reduced holonomy is one of the following groups: $SO(n)$, $U(\frac{n}{2})$ $(n \geq 4)$, $SU(\frac{n}{2})$ $(n \geq 4)$, $Sp(1) \cdot Sp(\frac{n}{4})$ $(n \geq 8)$, $Sp(\frac{n}{4})$ $(n \geq 4)$ and two exceptional groups $Spin(7)$ and G_2. We call (M, g) a Riemannian manifold with special holonomy if it is irreducible and its (reduced) holonomy is strictly contained in $SO(n)$. It can be shown that a Riemannian manifold of special holonomy is automatically Einstein if its holonomy is other than $U(\frac{n}{2})$. We have Kähler-Einstein

metrics for the $U(\frac{n}{2})$ case. In fact, these special Einstein metrics are of self-dual type. Each manifold (M,g) of special holonomy has a parallel $n-4$ form defined as follows: Let $W \subset \Lambda^2 TM$ be the subspace generated by the Lie algebra of the holonomy group of (M,g), then the curvatures lie in $S^2(W)$. Define a 4-form $\psi(W)$ by

$$\psi(W) = \sum w_i \wedge w_i, \tag{1.6}$$

where $\{w_i\}$ is an orthonormal basis of W. Clearly, it is independent of the choice of $\{w_i\}$ and is parallel. This 4-form induces a symmetric operator $T_{\psi(W)} : W \mapsto W$: $T_{\psi(W)}(v) = i_v\psi(W)$, where i_v denotes the interior product with v. One can check that $T_{\psi(W)}$ has at most two distinct eigenvalues. Moreover, there is a distinguished eigenspace W_0 of $T_{\psi(W)}$ of codimension 0, 1 and 3. Let β be the corresponding eigenvalue. Put

$$\Omega(W) = \frac{1}{\beta} * \psi(W). \tag{1.7}$$

Clearly, it is parallel. Denote by $S^2W = S^2W_0 + S^2W_1$ the decomposition according to eigenvalues, then the curvature $R(g)$ of g, which lies in S^2W, is decomposed into $R_0 \in S^2W_0$ and $R_1 \in S^2W_1$. Furthermore, we have

$$R_0 \wedge \Omega = *R_0 \tag{1.8}$$

and R_1, which can be void, is completely determined by Ricci curvature of g. Therefore, manifolds of special holonomy are always self-dual. It is very important in the study of Einstein metrics of special holonomy. For example, the self-duality implies an a prior L^2-bound on curvature: There is a uniform constant $C(p_1(M), \Omega, s(g))$, depending only on the first Pontrjagin class, Ω and the scalar curvature $s(g)$, such that for any Einstein metric g of special holonomy, we have

$$\int_M |R(g)|^2 dv = C(p_1(M), \Omega, s(g)). \tag{1.9}$$

Here Ω is the corresponding parallel form. In dimension 4, we can study self-dual Einstein metrics, that is, Einstein metrics with self-dual Weyl curvature. These self-dual metrics share similar properties as those with special holonomy do.

The special geometry we see most is the Kähler geometry. A Kähler manifold is a Riemannian manifold (M,g) whose holonomy lies in $U(\frac{n}{2})$, it is equivalent to saying that M has a compatible and parallel complex structure J, that is, $g(Ju,Jv) = g(u,v)$, where $u,v \in TM$ are arbitrary, and $\nabla J = 0$. So M is a complex manifold with induced complex structure by J. Usually, we denote g by its Kähler form $\omega_g = g(\cdot, J\cdot)$. In local complex coordinates $z_1, \cdots, z_m$ of M $(n = 2m)$,

$$\omega_g = \frac{\sqrt{-1}}{2} \sum_{i,j=1}^{m} g_{i\bar{j}} dz_i \wedge d\bar{z}_j, \tag{1.10}$$

where $(g_{i\bar{j}})$ is a positive Hermitian matrix-valued function. The self-duality simply means that the curvature of a Kähler metric has only components of type (1,1). A Kähler metric g is Einstein if and only if the trace of its curvature against ω_g is

constant, we call such a metric Kähler-Einstein. A necessary condition for the existence of Kähler-Einstein metric on M is that the first Chern class $c_1(M)$ is definite. Since the Ricci curvature of a Kähler metric g can be expressed as $-\partial\overline{\partial}\log det(g_{i\bar{j}})$, the Einstein equation is reduced to solving the following complex Monge-Amperé equation

$$\det(g_{i\bar{j}} + \frac{\partial^2\varphi}{\partial z_i \partial \overline{z}_j}) = e^{h-\lambda\varphi}\det(g_{i\bar{j}}), \qquad (g_{i\bar{j}} + \frac{\partial^2\varphi}{\partial z_i \partial \overline{z}_j}) > 0, \tag{1.11}$$

where φ is unknown and h is a given function depending only on g. This is a fully nonlinear elliptic equation and easier to solve.

A program initiated by E. Calabi in early 1950's is to study the existence and uniqueness of Kähler-Einstein metrics.[1] The uniqueness of Kähler-Einstein metrics was known in 1950's in the case that the first Chern class is nonpositive and was proved by Bando-Mabuchi [BM] in 1986 in the case that the first Chern class is positive. The difficult part of Calabi's program is about the existence. The celebrated solution of Yau [Ya] for the Calabi conjecture established the existence of a Ricci-flat metric, now named as Calabi-Yau metric, in each Kähler class on a compact Kähler manifold M with $c_1(M) = 0$. If $c_1(M) < 0$, the existence of Kähler-Einstein metrics was proved by Yau [Ya] and Aubin [Au], independently. There are further analytic obstructions to the existence of Kähler-Einstein metrics on M with $c_1(M) > 0$. Matsushima proved that M has a Kähler-Einstein metric only if the Lie algebra $\eta(M)$ of its holomorphic fields is reductive. Also if M has a Kähler-Einstein metric, then the Futaki invariant from [Fu] vanishes. The Futaki invariant is a character of $\eta(M)$. If M is a complex surface with $c_1(M) > 0$, then it admits a Kähler-Einstein metric if and only if the Lie algebra of holomorphic vector fields is reductive [Ti1]. For a general M with $c_1(M) > 0$, the existence of Kähler-Einstein metrics is equivalent to certain analytic stability [Ti2]. This analytic stability amounts to checking an nonlinear inequality of Moser-Trudinger type: Assume that $\eta(M) = \{0\}$. [2] If ω is a Kähler metric with $[\omega] = c_1(M)$ and φ with $\int_M \varphi\omega^n = 0$ and $\omega + \partial\overline{\partial}\varphi > 0$,

$$\log\left(\int_M e^{-\varphi}\omega^n\right) \leq J_\omega(\varphi) - f(J_\omega(\varphi)), \tag{1.12}$$

where f is some function bounded from below and satisfies $\lim_{t\to\infty} f(t) = \infty$[3] and J_ω is defined by

$$J_\omega(\varphi) = \sum_{i=0}^{n-1} \frac{i+1}{n+1}\frac{\sqrt{-1}}{2V}\int_M \partial\varphi\wedge\overline{\partial}\varphi\omega^i\wedge(\omega+\partial\overline{\partial}\varphi)^{n-i-1}, \tag{1.13}$$

where $V = \int_M \omega^n$. The inequality (1.12) has been checked for many manifolds, such as Fermat hypersurfaces. Furthermore, the analytic stability implies the asymptotic

[1] Later in 1980's, E. Calabi extended this to extremal Kähler metrics, one can see X.X. Chen's paper in this proceeding for recent progresses on extremal metrics.

[2] If $\eta(M) \neq \{0\}$, then the inequality holds only for those functions perpendicular to functions induced by holomorphic vector fields.

[3] f may depend on ω.

CM-stability of M introduced in [Ti2] in terms of Geometric Invariant Theory. If one proved the partial C^0-estimate conjectured in [Ti3], this asymptotic stability in [Ti2] would imply the existence of Kähler-Einstein metrics. Very recently, by using the Tian-Yau-Zeldich expansion (cf. [Ti4], [Cat], [Zel]) and a result of Zhiqin Lu [Lu], S. Donaldson [Do1] proved the asymptotic Chow stability [Mu] of algebraic manifolds which admit Kähler-Einstein metrics [Do1]. This gives a partial answer to one conjecture of Yau: If $\eta(M) = 0$, then there is a Kähler-Einstein metric on M if and only if M is asymptotically Chow stable. It would be a very interesting problem in algebraic geometry to compare the Chow stability with the CM-stability introduced in [Ti2]. Both stabilities can be defined in terms of the Chow coordinate of M, but their corrsponding polarizations are different (cf. [Paul]).

Kähler-Ricci solitons arose naturally from the study of the existence of Kähler-Einstein metrics and Hamilton's Ricci flow in Kähler geometry and generalize Kähler-Einstein metrics. A Kähler metric g is a Kähler-Ricci soliton if there is a holomorphic field X such that

$$\mathrm{Ric}(g) - \lambda\omega_g = L_X\omega_g. \tag{1.14}$$

As before, this equation can be reduced to a sightly more complicated complex Monge-Ampere equation (cf.[TZ]). It was proved in [TZ] that Kähler-Ricci solitons are unique modulo automorphisms. In subsequent papers, we also gave an analytic criterion for the existence as one did in [Ti2]. It was conjectured that given any Kähler manifold M with $c_1(M) > 0$, either M has a Kähler-Einstein metric or there are diffeomorphisms ϕ_i and Kähler metrics g_i such that $\phi_i^* g_i$ converge to a unique Kähler-Ricci soliton on M', which may be different from M. This conjecture was posed by R. Hamilton in studying the Ricci flow and myself in studying Kähler-Einstein metrics. When the complex dimension of M is 2, in view of the main result in [Ti1], it suffices to show that the blow-up of $\mathbb{C}P^2$ at two points admits a Kähler-Ricci soliton. This should be doable.

So far, the most successful method in proving the existence is the continuity method. The other possible approach is to use the Kähler-Ricci flow, which has only partial success (cf. X.X.Chen's talk at this ICM).

There remain many problems in studying Kähler-Einstein with prescribed singularities, though a lot has been done (cf. [CY], [TY1], [Ts], etc.). A given Kähler manifold M may not have definite first Chern class, so it does not admit any Kähler-Einstein metrics, but by blowing down certain subvarieties, the resulting manifold (possibly singular) may admit a canonical Kähler-Einstein metric. For instance, if M is an algebraic manifold of general type, can any given Kähler metric be deformed along the Kähler-Ricci flow to a unique Kähler-Einstein metric? Is the limiting metric independent of the initial metric? The answer to these questions seems to be affirmative in complex dimension 2 or in the case that minimal models exist. Another unsolved problem is Yau's conjecture: Every complete Calabi-Yau open manifold can be compactified such that the divisor at infinity is the zero-locus of a section of a line bundle proportional to the anti-canonical bundle. This is a hard problem. In [TY2], [TY3] and [BK], complete Calabi-Yau manifolds were constructed on complements of a smooth divisor which is a fraction of anti-canonical

divisor and satisfies certain positivity conditions (also see [Jo]). In view of these and [CT1], one is led to the following conjecture: a complete Calabi-Yau manifold M with quadratic curvature decay and euclidean volume growth is of the form $M = \overline{M} \backslash D$ such that D is ample near D and the anti-canonical bundle $K_{\overline{M}}^{-1}$ is $\alpha[D]$ for some $\alpha > 1$. This can be considered as the refinement of Yau's conjecture in a special case.

The next special holonomy is contained in $Sp(1)Sp(\frac{n}{4})$. Riemannian manifolds with such a holonomy are called quaternion-Kähler manifolds. They are automatically Einstein. The prototype is the quaternionic projective space $\mathbf{P}_{\mathbf{H}}^{\frac{n}{4}}$. There are many examples of quaternion-Kähler manifolds due to the works of many people, including S. Salamon, Galicki-Lawson, Lebrun, etc. Quaternion-Kähler manifolds with zero scalar curvature are hyperKähler, that is, its holonomy lies in $Sp(\frac{n}{4})$. The existence of hyperKähler manifolds follows from Yau's solution for the Calabi conjecture. However, we do not know yet if there are quaternion-Kähler manifolds with positive scalar curvature and which are not locally symmetric, while we do have a number of symmetric ones, the so-called Wolf spaces. It led Lebrun and S. Salamon to guess that the Wolf spaces are all complete quaternion-Kähler manifolds with positive scalar curvature. So far, it has been checked up to dimension $\frac{n}{4} \leq 3$. We would like to point out that there are many non-symmetric quaternion-Kähler orbifolds with positive scalar curvature due to Galicki-Lawson ([GL]).

Riemannian manifolds with holonomy G_2 and Spin(7) must be Ricci-flat and of dimension 7 and 8, respectively. It took a long time to settle the question of whether such metrics exist, even locally. Local metrics with these holonomy were constructed by R. Bryant [Br]. Later, complete examples were constructed by Bryant and S. M. Salamon[BS]. Examples of compact 7- and 8-manifolds with holonomy G_2 and Spin(7) were first constructed by D. Joyce in early 1990's (cf. [Jo]). D. Joyce's construction was inspired by the Kummer construction: metrics with holonomy SU(2) on the $K3$ surface can be obtained by resolving the 16 singularities of the orbifolds T^4/Z_2, where Z_2 acts on T^4 with 16 fixed points. In the case of G_2, Joyce chooses a finite group $\Gamma \subset G_2$ of automorphisms of the torus T^7. Then he resolves the singularities of T^7/Γ to get a compact 7-manifold M with holonomy G_2. A similar construction can be implemented for the Spin(7) case by choosing a finite group Γ of automorphisms of the torus T^8 and a flat Γ-invariant Spin(7)-structure on T^8. More recently, Kovalev gave a new construction of Riemannian metrics with special holonomy G_2 on compact 7-manifolds. The construction is based on gluing asymptotically cylindrical Calabi-Yau manifolds built up on the work in [TY2]. Examples of new topological types of compact 7-manifolds with holonomy G_2 were obtained.

So far, all Ricci-flat compact manifolds are of special holonomy. There should exist complete Ricci-flat manifolds with generic holonomy $\mathrm{SO}(n)$. The question is how we can find them. Here is a possible example in 4-dimension: It was shown that there is a Calabi-Yau manifold with cylindrical end asymptotic to T^3 [TY2]. Complex analytically, this manifold can be obtained by blowing up the 9 base points of a generic elliptic pencil on $\mathbb{C}P^2$ and removing one smooth fiber. Now take two copies of such Calabi-Yau manifolds and glue them along the T^3's at infinity. One

way of gluing them is to respect the complex structures, then we will get a $K3$ surface. Could one use different gluing maps which do not preserve the complex structures, so that one may obtain new Ricci-flat manifolds with generic holonomy? We can also ask if any complete Ricci-flat manifolds can be decomposed in some sense into a connected sum of Calabi-Yau manifolds. Similar things can be done in higher dimensions.

Geometry of moduli space of Einstein manifolds is extremely important. For example, the moduli space of Calabi-Yau manifolds provides the B-model in the Mirror Symmetry. If (M, g) is an Einstein manifold with special holonomy, then it was proved that nearby Einstein manifolds in the moduli space is also of special holonomy. The first analytic problem about the moduli is its compactness. The moduli space is very often noncompact, so we need to compactify it. Then we can consider what structures a compactified moduli space has.

We have pointed out before that for any Einstein manifold (M, g) with special holonomy, the L^2-norm of its curvature depends only on the second Chern character and the Einstein constant λ (assuming that the volume of M is normalized, say 1). One can first give a weak compactification $\overline{\mathcal{M}}$ of the moduli space $\mathcal{M}$ of Einstein manifolds with special holonomy in the Gromov-Hausdorff topology. A basic problem is the regularity of a limit in $\overline{\mathcal{M}}\backslash\mathcal{M}$. There are two cases of the limit, one is when the limit is still compact, while the other has infinite diameter as a length space. Here let us consider only the first case, since we know much more in this case and it is necessary for studying the second. If M_∞ is a compact limit, then there is a sequence of Einstein manifolds (M_i, g_i) with special holonomy and bounded diameter converging to M_∞ in the Gromov-Hausdorff topology. When the dimension is 2, it was proved in [Ti1] that M_∞ is a Kähler-Einstein orbifold with isolated singularities. Its real version was done by M. Anderson in [An]. The compactness theorem played a very important role in the resolution of the Calabi problem for complex surfaces (cf. [Ti1]). In [CT2], Cheeger and I proved

Theorem 1.1. [4] *Let M_∞ be the above limit of a sequence of Einstein manifolds (M_i, g_i) with the same special holonomy and uniformly bounded diameter. Then there is a rectifiable closed subset $S \subset M_\infty$ such that $M_\infty\backslash S$ is a smooth manifold which admits an Einstein metric g_∞ with the same holonomy as (M_i, g_i) do. Furthermore, M_∞ is the metric completion of $M_\infty\backslash S$ with respect to the distance induced by g_∞.*

This is based on deep works of Cheeger-Colding [CC] on spaces which are limits of manifolds with Ricci curvature bounded from below, my joint work with Cheeger and Colding on structure of the singular sets of limits of Einstein manifolds with L^2 curvature bounds [CCT] and Cheeger's work on rectifiability of singular sets of the limits [Che]).

Remark 1.2. In fact, the convergence can also be strengthened: There is an exhausion of $M_\infty\backslash S$ by compact sets $K_i \subset K_{i+1} \cdots$ and diffeomorphisms $\phi_i : K_i \mapsto M_i$ such that $\phi_i(K_i)$ converge to S in the Gromov-Hausdorff topology and $\phi^* g_i$ converge to g_∞ in the C^∞-topology.

[4] The Kähler case of this theorem was proved in [CCT].

The most fundamental problem left is the regularity of S or structure of M_∞ along S. The conjecture is that S can be stratified into $\coprod_{i\leq n-4} S_i$ such that each stratum S_i is a smooth manifold of dimension $\leq i$. If (M_i, g_i) are Kähler-Einstein, then $S_{2j+1} = S_{2j}$. We know (cf. [CCT], [CT2]) that tangent cones at almost all points of S_{n-4} are of the form $\mathbb{R}^{n-4} \times C(S^3/\Gamma)$, where $C(S^3/\Gamma)$ is the cone over S^3/Γ and $\Gamma \subset SO(4)$ is a finite group. It makes us conjecture that M_∞ should be homeomorphic to an open set of $\mathbb{R}^{n-4} \times C(S^3/\Gamma)$ locally along $S_{n-4} \backslash \coprod_{i\leq n-5} S_i$. It is plausible that M_∞ is actually smooth along $S_{n-4} \backslash \coprod_{i\leq n-5} S_i$ in a suitable sense. We also believe that the $(n-4)$-form Ω_∞ associated to the special holonomy of g_∞ extends to S in a suitable sense and its restriction to S is the same as the volume, i.e., S is calibrated by Ω_∞ in a suitable sense.

When (M_i, g_i) are Kähler-Einstein manifolds with positive scalar curvature, it was conjectured by the author long time ago that a multiple of the anticanonical bundle of $M_\infty \backslash S$ extends to be a line bundle across the singular set S. This is of course true if one can show that M_∞ has only quotient singularities. The affirmation of this conjecture will enable us to prove the converse of a result in [Ti2], that is, the algebraic stability of a Kähler manifold with positive first Chern class assures the existence of Kähler-Einstein metrics.

Finally, we shall refer the readers to [CT2] for detailed study of tangent cones at any singularity of M_∞.

2. Yang-Mills equation

Next we discuss the Yang-Mills equation. The Yang-Mills equation has played a fundamental role in our study of physics and geometry and topology in last few decades. In the following, unless specified, we assume that (M, g) is a Riemannian manifold of dimension n and $\mathbf{G}$ is a compact subgroup in $\mathbf{SO}(r)$ and $\mathbf{g}$ is its Lie algebra. Let E be a $\mathbf{G}$-bundle over M.

First we recall that a connection of E over M is locally of the form

$$A = A_i dx_i, \quad A_i \in \mathbf{g} \tag{2.1}$$

where $x_1, \cdots, x_n$ are euclidean coordinates of $\mathbb{R}^n$ and A_i are matrices in $\mathbf{g}$. Its curvature can be computed as follows:

$$F_A = dA + A \wedge A. \tag{2.2}$$

The Yang-Mills functional is defined on the space of connections and given by

$$\mathcal{Y}(A) = \frac{1}{4\pi^2} \int_M |F_A|_g^2 dV_g. \tag{2.3}$$

The Yang-Mills equation is simply its Euler-Lagrange equation

$$D_A^* F_A = 0, \tag{2.4}$$

where D_A denotes the covariant derivative of A and D_A^* is its adjoint. On the other hand, being the curvature of a connection, A automatically satisfies the second

Bianchi identity $D_A F_A = 0$. We will call A a Yang-Mills connection if it satisfies (2.4).

The gauge group $\mathcal{G}$ consists of all sections of $Ad(E)$ over M, locally, they are just maps into $\mathbf{G} \subset \mathbf{SO}(r)$. It acts on the space of connections by assigning A to $\sigma(A) = \sigma A \sigma^{-1} - \sigma d\sigma^{-1}$ for each $\sigma \in \mathcal{G}$. Clearly, the Yang-Mills functional is invariant under the action of $\mathcal{G}$, so does the Yang-Mills equation. In particular, it implies that the Yang-Mills equation is not elliptic. A difficult problem is to construct good gauges which can be controled by curvatures. So called Coulomb gauges have been constructed by Uhlenbeck [Uh2] in $L^{n/2}$-norms and more recently, by Tao-Tian and Meyer-Riviere in Morrey norms (cf. [TT]).

The simplest Yang-Mills connections are provided by harmonic one forms: If $G = U(1)$, then $\mathbf{g} = i\mathbb{R}$ and A is simply a one-form and the Yang-Mills equation is $d^* dA = 0$, the gauge transformation is given by $\sigma = e^{ia} \mapsto A + ida$. It follows that modulo gauge transformations, abelian Yang-Mills connections are in one-to-one correspondence with harmonic one forms.

Now we assume that (M, g) is of special holonomy. Let Ω be the associated closed form of degree $n-4$. We say that a connection A is Ω-self-dual if

$$*(\Omega \wedge F_A) = F_A, \tag{2.5}$$

where $*$ is the Hodge operator [5]

One can show that an Ω-self-dual connection is a Yang-Mills connection. Clearly, the self-duality is invariant under gauge transformations. So self-dual connections provide a special class of Yang-Mills solutions.

There are many examples of Ω-self-dual connections. First, the Levi-Civita connection of the underlying Riemannian metric is Ω-self-dual. In this sense, the Yang-Mills equation is a semi-linear version of the Einstein equation. Secondly, if E is a stable holomorphic vector bundle, then the Donaldson-Yau-Uhlenbeck theorem ([Do2], [UY]) states that E has a unique Hermitian-Yang-Mills connection, an easy computation shows that a connection is Hermitian-Yang-Mills if and only if it is Ω-self-dual, where $\Omega = -\frac{\omega^{n/2-2}}{(n/2-2)!}$ and ω is the underlying Kähler form. Thirdly, if (M, g) is a Calabi-Yau 4-fold and Ω is its associated $(n-4)$-form induced by the $SU(4) \subset Spin(7)$- structure, then Ω-self-dual connections are just complex self-dual instantons of Donaldson-Thomas [DT]. Also, one may construct Ω-self-dual instantons from Ω-calibrated submanifolds of M.

When $n = 4$, self-dual instantons were used to construct the Donaldson invariants for 4-manifolds. This eventually led to the Seiberg-Witten invariants, which is much easier to compute. The construction goes roughly as follows: Let M be a 4-manifold and g is a generic metric. Let E be an $SU(2)$-bundle over M. Consider the moduli space $\mathcal{M}_E$ of self-dual instantons of E, that is, solutions of

$$F_A = *F_A \tag{2.6}$$

[5] If $\mathbf{G} \subset U(r)$, one can consider more general self-dual equation: A is an Ω-self-dual connection if $\mathrm{tr}(F_A)$ is harmonic and $*(\Omega \wedge F_A^0) = F_A^0$, where $F_A^0 = F_A - \frac{1}{r}\mathrm{tr}(F_A)Id$. If $\mathbf{G} = \mathrm{SU}(r)$, this coincides with (2.5).

modulo gauge transformations. A generalized instanton consists of an anti-self-dual instanton and a tuple of points of M such that the second Chern class of the instanton and totality of the points sum up to represent $c_2(E)$. Let $\overline{\mathcal{M}}_E$ be the moduli space of all generalized instantons of E. Then Uhlenbeck compactness theorem states that $\overline{\mathcal{M}}_E$ is compact. Also $\overline{\mathcal{M}}_E$ is a stratified space with $\mathcal{M}_E$ as its main stratum. If $b_2^+(M) \geq 3$ and g is generic, then each stratum has expected dimension which can be easily computed by the Atiyah-Singer index theorem, so $\overline{\mathcal{M}}_E$ can be taken as a fundamental class. The Donaldson invariants are obtained by integrating pull-backs of cohomology classes of M on this fundamental class.

Similarly, one can define the Seiberg-Witten invariant by using the Seiberg-Witten equation. Technically, it is much easier since the moduli space is already compact. Sometimes, it was said that the Seiberg-Witten invariant does not need hard analysis, in fact, it is false. Taubes' deep theorem on equivalence of Seiberg-Witten and Gromov-Witten invariants requires hard analysis.

What about higher dimensional cases? Can we construct new deformation invariants by using Ω-self-dual connections? In order to achieve it, one has to consider the following issues: 1. Is the corresponding self-dual equation elliptic? Indeed, the self-dual equation on a $Spin(7)$-manifold is elliptic. 2. Can we compactify the moduli space? If so, how do we stratify the compactified moduli space? 3. Does each stratum have right dimension which can be predicted by the index theorem? If we solve these issues, we can define new invariants, then we can study how to compute them.

The first issue is easy to check. We just need to linearize the self-dual equation and see if it is elliptic. There are examples, such as, self-dual equations on 4-manifolds and Donaldson-Thomas complex self-dual on Calabi-Yau 4-manifolds. It will be very useful to construct deformation invariants by using complex self-dual instantons. The success of it will provide a powerful tool of constructing holomorphic cycles of codimension 4, which are pretty much evading us.

Next we consider the compactification. Having a good compactification, we will be able to get property 3 in the above. Let (M, g) be a compact Riemannian manifold of dimension n and with special holonomy. Let Ω be the associated closed form of degree $n-4$. Let E be a unitary vector bundle over M. Recall that $\mathfrak{M}_{\Omega,E}$ consists of all gauge equivalence classes of Ω-asd instantons of E over M. In general, $\mathfrak{M}_{\Omega,E}$ may not be compact. So we will compactify it.

An admissible Ω-self-dual instanton is simply a smooth connection A of E over $M\backslash S(A)$ for a closed subset $S(A)$ of Hausdorff dimension $n-4$ such that $\int_M |F_A|^2 < \infty$. A generalized Ω-self-dual instanton is made of an admissible Ω-self-dual instanton A of E and a closed integral current $C = C_2(S, \Theta)$ calibrated by Ω, such that cohomologically,

$$[C_2(A)] + \mathrm{PD}[C_2(S, \Theta)] = C_2(E), \tag{2.7}$$

where $C_2(A)$ denotes the Chern-Weil form of A[6] and $C_2(E)$ denotes the second Chern class of E. Two generalized Ω-self-dual instantons (A, C), (A', C') are equiv-

[6]One can show this form, which was originally defined on $M\backslash S(A)$, extends to a well-defined current on M.

alent if and only if $C = C'$ and there is a gauge transformation σ on $M \backslash S(A) \cup S(A')$, such that $\sigma(A) = A'$ on $M \backslash S(A) \cup S(A')$. We denote by $[A, C]$ the gauge equivalence class of (A, C). We identify $[A, 0]$ with $[A]$ in $\mathcal{M}_{\Omega,E}$ if A extends to a smooth connection of E over M modulo a gauge transformation. We define $\overline{\mathfrak{M}}_{\Omega,E}$ to be set of all gauge equivalence classes of generalized Ω-self-dual instantons of E over M.

The topology of $\overline{\mathfrak{M}}_{\Omega,E}$ can be defined as follows: a sequence $[A_i, C_i]$ converges to $[A, C]$ in $\overline{\mathfrak{M}}_{\Omega,E}$ if and only if there are representatives (A_i, C_i) such that their associated currents $C_2(A_i, C_i)$ converge weakly to $C_2(A, C)$ as currents, where

$$C_2(A', C') = C_2(A') + C_2(S', \Theta'), \quad C' = (S', \Theta'). \tag{2.8}$$

It is not hard to show that by taking a subsequence if necessary, $\tau_i(A_i)$ converges to A outside $S(A)$ and the support of C for some gauge transformations τ_i.

The following was proved in [Ti5] and provides a compactification for the moduli space of Ω-self-dual connections.

Theorem 2.1. *For any M, g, Ω and E as above, $\overline{\mathfrak{M}}_{\Omega,E}$ is compact with respect to this topology.*

Of course, $\overline{\mathfrak{M}}_{\Omega,E}$ admits a natural stratification. The remaining problem, which is also important for issue 3, is about regularity of a generalized Ω-self-dual instanton. Another interesting problem is to develop a deformation theory of smoothing singular self-dual instantons. Are there any constraints on a singular self-dual instanton which is the limit of smooth self-dual instantons? We do not even know any example of a Hermitian-Yang-Mills connection with an isolated singularity and which can be approximated by smooth Hermitian-Yang-Mills connections.

Let us give an example. Assume that (M, g) is a Kähler manifold with Kähler form ω. Put $\Omega = \omega^{m-2}/(m-2)!$, where $n = 2m$. Then an Ω-self-dual instanton A is simply a Hermitian-Yang-Mills connection, that is $F_A^{0,2} = 0$ and $F_A^{1,1} \cdot \omega = 0$, where $F_A^{k,l}$ is the (k, l)-part of F_A. If (A, C) is a generalized Ω-self-dual instanton, it follows from a result of J. King that there are positive integers m_a and irreducible complex subvarieties V_a such that for any smooth φ with compact support in M,

$$C_2(S, \Theta)(\varphi) = \sum_a m_a \int_{V_a} \varphi.$$

On the other hand, using a result of Bando-Siu, one can show [TYa] that there is a gauge transformation τ such that $\tau(A)$ extends to be a smooth connection outside a complex subvariety of codimension greater than 2.

We expect that general self-dual connections have analogous properties. If (A, C) is a generalized Ω-self-dual connection, we would like to have (1) the regularity of the current C, that is, C is presented by finitely many Ω-calibrated subvarieties with integral multiplicity; (2) There is a gauge transformation σ such that $\sigma(A)$ extends to a smooth connection outside a subvariety of codimension at least 6. In [Uh2], any isolated singularity of a Yang-Mills connection in dimension 4 can be removed. In [TT], a removable singularity theorem was established for stationary Yang-Mills connections in higher dimensions. Using this, we can conclude

that $\sigma(A)$ extends to a smooth connection outside a closed subset S with vanishing $(n-4)$-Hausdorff measure. Further understanding on S is needed. We will discuss regularity of C in the next section.

A particularly interesting case is the complex self-dual instanton. We do expect to construct new invariants for Calabi-Yau 4-folds by showing that the above moduli space of generalized complex self-dual instantons gives rise to a fundamental class. The main problem left is the regularity of generalized instantons. A special case of this can be done nicely. If the underlying Calabi-Yau 4-fold M is of the form $Y \times T^1_{\mathbb{C}}$, where Y is a Calabi-Yau 3-fold and $T^1_{\mathbb{C}}$ is a complex 1-torus, then a $T^1_{\mathbb{C}}$-invariant complex self-dual instanton is given by a Hermitian Yang-Mill connection A on Y and a $(0,3)$-form f with $\overline{\partial}^* f = 0$. The expected dimension of its moduli space is zero. Counting them with sign gives rise to the holomorphic Casson invariant, which was constructed previously by R. Thomas using the virtual moduli cycle construction in algebraic geometry[Th].

Other analytic problems on the Yang-Mills equation include whether or not the Yang-Mills flow develops singularity at finite time. It was proved by Donaldson that the Yang-Mills flow along Hermitian metrics of a holomorphic bundle has global solution. Of course, if the dimension of the underlying manifold is less than 4, the Yang-Mills flow has a global solution. In general, it is still open. If a singularity forms at finite time, how does it look like?

3. Minimal submanifolds

The study of minimal submanifolds is a classical topic. We will not intend to cover all aspects of this topic. We will only discuss issues related to previous discussions and particularly self-dual type solutions of the minimal submanifold equation.

Let (M, g) be an n-dimensional Riemannian manifold and S be a submanifold in M. Recall that S is minimal if its mean curvature H_S vanishes. The mean curvature arises from the first variation of volume of submanifolds. Minimal submanifolds are closely related to the Yang-Mills equation. In fact, it was shown in [Ti5] that if a Yang-Mills connection has its curvature concentrated along a submanifold, then this submanifold must be minimal and of codimension 4. Motivated by this, recently, S. Brendle, etc. developed a deformation theory of constructing Yang-Mills connections from minimal submanifolds.

Now assume that M has a closed differential form Ω with its norm $|\Omega| \leq 1$. A submanifold S is calibrated by Ω if $\Omega|_S$ coincides with the induced volume form. Calibrated submanifolds are minimal (cf. [HL]). The study of calibrated submanifolds was pioneered in the seminal work [HL] of Harvey and Lawson. It now becomes extremely important in the string theory. As we have seen in the above, they also appear in formation of singularity in the Einstein equation. In particular, when a self-dual connection has its curvature concentrated along a submanifold, this submanifold is calibrated [Ti5], so a calibrated submanifold can be regarded as a self-dual solution of the minimal submanifold equation.

Let (M, ω) be a symplectic manifold and J be a compatible almost complex

structure, that is, $\omega(Ju, Jv) = \omega(u, v)$ and $\omega(u, Ju) > 0$ for any non-zero tangent vectors u and v. Define a compatible metric g by $g(u, v) = \omega(u, Jv)$. Any ω-calibrated submanifolds are J-holomorphic curves, that is, each tangent space is a J-invariant subspace in TM. They are particularly minimal with respect to g. Holomorphic curves have been used to establish a mathematical foundation of the quantum cohomology, the mirror symmetry, etc. (cf. [RT]). The key of it is to construct the Gromov-Witten invariants by showing the moduli space of J-holomorphic curves can be taken as a fundamental class in a suitable sense. This was proven by first constructing a "nice" compactification of the moduli space of J-holomorphic curves and then applying appropriate transversality theory.

I do believe that there should be new invariants by using other calibrated submanifolds. A particularly interesting case is the Cayley cycles in a Spin(7)-manifold. Note that a Calabi-Yau 4-fold is a special Spin(7)-manifold. Again the problem is about the structure of singular Cayley cycles. This proposed new invariants will provide a powerful tool of constructing Cayley cycles in a Spin(7)-manifold, particularly, holomorphic and special Lagrangian cycles in a Calabi-Yau 4-fold. A related problem is to construct new invariants for hyperKähler manifolds by using tri-holomorphic maps. A good compactification for the moduli of tri-holomorphic maps is needed, but this should be technically easier. Partial results have been obtained in [LiT] and [CL1].

Another possible invariant may exist for Calabi-Yau manifolds. Let (M, ω) be a Calabi-Yau n-fold with a holomorphic n-form ω such that

$$\omega^n (-1)^{\frac{n(n-1)}{2}} n! \left(\frac{\sqrt{-1}}{2}\right)^n \Omega \wedge \overline{\Omega}. \tag{3.1}$$

A special Lagrangian submanifold is a submanifold $L \subset M$ such that $\omega|_L = 0$ and Ω restricts to the induced volume form of L. If one has a good compactification theorem for special Lagrangian submanifolds, then one can count them to obtain a new invariant for M. A particularly important case is for Calabi-Yau 3-folds.

The minimal equation is nonlinear and does have singular solutions. So one has to introduce weak solutions. An integral k-dimensional current $C = (S, \Theta, \xi)$ consists of a k-dimensional rectifiable set S [7] of locally finite Hausdorff measure, an H^k-integrable integer-valued function Θ and a k-form $\xi \in \wedge^k TS$ with unit norm. Each current induces a natural functional Φ_C on smooth forms with compact support: For any smooth form φ,

$$\Phi_C(\varphi) = \int_S \langle \varphi, \xi \rangle dH, \tag{3.2}$$

where dH^k denotes the k-dimensional Hausdorff measure. We say C has no boundary if $\Phi_C(d\psi) = 0$ for any ψ. One can define the generalized mean curvature of C as the variation of volume. A current C is minimal if its mean curvature vanishes. A current C is calibrated by a k-form Ω if $\Omega|_{T_x S}$ coincides with the induced volume form whenever the tangent space $T_x S$ exists. Of course, a calibrated current is minimal, provided that $d\Omega = 0$ and $|\Omega| \leq 1$.

[7] This implies that there is a unique tangent space at a.e. point of S.

A fundamental problem in the regularity theory of minimal surfaces is the regularity of minimizing currents. A result of F. Almgren claims that an area minimizing current is regular outside a subset of Hausdorff codimension two [Alm]. In many geometric applications, we will encounter with calibrated currents, for example, in the famous work of Taubes on equivalence of the Seiberg-Witten invariants and the Gromov invariants, the key technical point is to show that any ω-calibrated current in a symplectic 4-manifold (M, ω) is a classical minimal surface, i.e., the image of a pseudo-holomorphic map from a smooth Riemann surface ([Ta], also see [RiT]).[8] This current is obtained as an adiabatic limit of curvature forms of solutions of deformed Seiberg-Witten equations. The problems of this type should also occur when we study the Calabi-Yau manifolds near large complex limits. Of course, this regularity problem also appears in compactifying moduli spaces of calibrated cycles.

Here is what we think should be true: If $C = (S, \Theta, \xi)$ is a k-dimensional calibrated current, then S can be stratified into $\coprod_i S_i$ such that each stratum S_i is a smooth manifold of dimension i, which is at most $k-2$, and Θ is constant on each stratum.

If the calibrating form Ω is $\omega^l/l!$ $(k = 2l)$ on a symplectic manifold (M, ω) with a compatible metric g, then Ω-calibrated current is pseudo-holomorphic, that is, any tangent space is invariant under the almost complex structure induced by ω and g. In this special case, the conjecture is that S is stratified into pseudo-holomorphic strata S_{2j}. If the dimension of C is 2, then the conjecture claims that C is induced by a pseudo-holomorphic curve. This conjecture follows from a result of King when (M, g) is Kähler, i.e., the corresponding almost complex structure is integrable. Very recently, Riviere and I can prove this conjecture when $\dim C = 2$. When $\dim M = 4$, it was already known (cf. [Ta], [RiT]).

A nice way of deforming a surface into a minimal one is to use the mean curvature flow. If (M, g) is a compact Kähler-Einstein surface and S_0 is a symplectic surface with respect to the Kähler form, then surfaces along the mean curvature flow starting from S_0 are also symplectic [ChT]. If the flow has a global solution, then S_0 can be deformed to a symplectic minimal surface. In particular, S_0 is isotopic to a symplectic minimal surface. However, it is highly nontrivial to show that the flow has a global solution. Partial results have been obtained ([CL2], [WaM]). Nevertheless, it was conjectured in [Ti6] that any symplectic surface in a Kähler-Einstein surface is isotopic to a symplectic minimal surface. This has been checked for a quite big class of symplectic surfaces in a Kähler-Einstein surfaces with positive scalar curvature (cf. [ST]) by using pseudo-holomorphic curves. One can also ask similar questions for the mean curvature flow along Lagrangian submanifolds. It was proved first by Smocsky [Sm] that the mean curvature flow preserves the Lagrangian property. We refer [ThY] for more discussions on the mean curvature flow for Lagrangian submanifolds.

[8]This is a special case of the main result in [SCh], which states that a 2-dimensional area minimizing is a classical minimal surface. But Chang's proof relies on some hard techniques of [Alm], so a self-contained proof is very desirable.

References

[Alm] F. Almgren, *Almgren's big regularity paper. Q-valued functions minimizing Dirichlet's integral and the regularity of area-minimizing rectifiable currents up to codimension 2*, World Scientific Monograph Series in Mathematics, 1. World Scientific Publishing Co., Inc., 2000.

[An] M. Anderson, *Ricci curvature bounds and Einstein metrics on compact manifolds*, J. Amer. Math. Soc., 2 (1989), no. 3, 455–490.

[Au] T. Aubin, *Some nonlinear problems in Riemannian geometry.* Springer, Berlin-Heidelberg-New York, (1997).

[BG] C. Boyer and K. Galicki, *3-Sasakian manifolds*, Surveys in differential geometry: Essays on Einstein manifolds, 123–184, Surv. Differ. Geom., VI, Int. Press, Boston, MA, 1999. S. Bando and R. Kobayashi, *Ricci-flat Khler metrics on affine algebraic manifolds*, Math. Ann., 287 (1990), no. 1, 175–180.

[BK] S. Bando and R. Kobayashi, *Ricci-flat Khler metrics on affine algebraic manifolds. II*, Math. Ann., 287 (1990), no. 1, 175–180.

[BM] S. Bando and T. Mabuchi, *Uniqueness of Einstein Kähler metrics modulo connected group actions*, Algebraic Geometry, Adv. Studies in Pure Math., 10 (1987).

[Br] R. Bryant, *Metrics with exceptional holonomy*, Ann. of Math., (2) 126 (1987), no. 3, 525–576.

[BS] R. Bryant and S. Salamon, *On the construction of some complete metrics with exceptional holonomy*, Duke Math. J., 58 (1989), no. 3, 829–850.

[Cat] D. Catlin, *The Bergman kernel and a theorem of Tian*, Analysis and geometry in several complex variables (Katata, 1997), 1–23, Trends Math., Birkhuser Boston, Boston, MA, 1999.

[CC] J. Cheeger and T. Colding, *On the structure of spaces with Ricci curvature bounded below, I-III*, J. Differential Geom., 46 (1997), no. 3, 406–480 and 54 (2000), no. 1, 13–74.

[CCT] J. Cheeger, T. Colding and G. Tian, *Constraints on singularities under Ricci curvature bounds*, C. R. Acad. Sci. Paris Sr. I Math. 324 (1997), no. 6, 645–649 (Its full version will appear in GAFA).

[Ch] B. Chow, *The Ricci flow on the 2-sphere*, J. Differential Geom., 33 (1991), no. 2, 325–334.

[Che] J. Cheeger, *L_p-bounds on curvature, elliptic estimates and rectifiability of singular sets*, C. R. Math. Acad. Sci. Paris 334 (2002), no. 3, 195–198.

[ChT] J.Y.Chen and G. Tian, *Moving symplectic curves in Khler-Einstein surfaces*, Acta Math. Sin. (Engl. Ser.), 16 (2000), no. 4, 541–548.

[CL1] J.Y. Chen and Jiayu Li, *Quaternionic maps between hyperkhler manifolds*, J. Differential Geom., 55 (2000), no. 2, 355–384.

[CL2] J.Y. Chen and Jiayu Li, *Mean curvature flow of surface in 4-manifolds*, Adv. Math., 163 (2001), no. 2, 287–309.

[CT1] J. Cheeger and G. Tian, *On the cone structure at infinity of Ricci flat manifolds with Euclidean volume growth and quadratic curvature decay*, Invent. Math., 118 (1994), no. 3, 493–571.

[CT2] J. Cheeger and G. Tian, *Compactifying moduli of special Einstein manifolds*,

in preparation.

[CY] S.Y. Cheng and S.T. Yau, *On the existence of a complete Khler metric on noncompact complex manifolds and the regularity of Fefferman's equation*, Comm. Pure Appl. Math., 33 (1980), no. 4, 507–544.

[Do1] S. Donaldson, *Scalar curvatures and projective imbeddings, I*, preprint, 2001.

[Do2] S. Donaldson, *Infinite determinants, stable bundles and curvature*, Duke Math. J., 54 (1987), no. 1, 231–247.

[DT] S. Donaldson and R. Thomas, *Gauge theory in higher dimensions*, The geometric universe (Oxford, 1996), 31–47, Oxford Univ. Press, Oxford, 1998.

[Fe] H. Federer, *Geometric measure theory.* Springer. Berlin-Heidelberg-New York, (1969).

[FO] K. Fukaya and K. Ono, *Arnold conjecture and Gromov-Witten invariants*, Topology 38 (1999), no. 5, 933–1048.

[Fu] A. Futaki, *An obstruction to the existence of Einstein-Kähler metrics*, Inv. Math., 73 (1983), 437–443.

[GL] K. Galicki and B. Lawson, *Quaternionic reduction and quaternionic orbifolds*, Math. Ann., 282 (1988), no. 1, 1–21.

[Ha] R. Hamilton, *The Ricci flow on surfaces*, Mathematics and general relativity (Santa Cruz, CA, 1986), 237–262, Contemp. Math., 71, Amer. Math. Soc., Providence, RI, 1988.

[HL] R. Harvey and H.B. Lawson, *Calibrated geometries*, Acta. Math., 148 (1982), 47–157.

[Jo] D. Joyce, *Compact manifolds with special holonomy.* Oxford Mathematical Monographs. Oxford University Press, Oxford, 2000.

[KW] J. Kazdan and F. Warner, *Existence and conformal deformation of metrics with prescribed Gaussian and scalar curvatures*, Ann. of Math. (2) 101 (1975), 317–331.

[LiT] Jiayu Li and G. Tian, *A blow-up formula for stationary harmonic maps*, Internat. Math. Res. Notices, 1998, no. 14, 735–755.

[LT] Jun Li and G. Tian, *Virtual moduli cycles and Gromov-Witten invariants of algebraic varieties*, J. Amer. Math. **11**(1998), no. 1, 119–174.

[Lu] Z.Q. Lu, *On the lower order terms of the asymptotic expansion of Tian-Yau-Zelditch*, Amer. J. Math., 122 (2000), no. 2, 235–273.

[Ma] Y. Matsushima, *Sur la structure du group homéomorphismes analytiques d'une certaine variété Kaehlérienne*, Nagoya Math. J., 11 (1957), 145–150.

[Mu] D. Mumford, *Stability of projective varieties*, Enseignement Math., (2) 23 (1977), no. 1-2, 39–110.

[Na] H. Nakajima, *Compactness of the moduli space of Yang-Mills connections in higher dimensions*, J. Math. Soc. Japan, 40 (1988).

[Paul] S. Paul and G. Tian, preprint, 2002.

[Pr] P. Price, *A monotonicity formula for Yang-Mills fields*, Manuscripta Math., 43 (1983), 131–166.

[RiT] T. Riviere and G. Tian, *The singular set of J-holomorphic maps into algebraic varieties*, preprint, 2001.

[RT] Y.B. Ruan and G. Tian, *A mathematical theory of quantum cohomology*, J.

Diff. Geom., 42 (1995), no. 2, 259–367.

[Ru] Y.B. Ruan, *Virtual neighborhoods and pseudo-holomorphic curves*, preprint, 1996.

[Sc] R. Schoen, *Conformal deformation of a Riemannian metric to constant scalar curvature*, J. Differential Geom., 20 (1984), no. 2, 479–495.

[SCh] S. Chang, *Two-dimensional area minimizing integral currents are classical minimal surfaces*, J. Amer. Math. Soc., 1 (1988), no. 4, 699–778.

[Sm] K. Smoczyk, *Harnack inequality for the Lagrangian mean curvature flow*, Calc. Var. Partial Differential Equations, 8 (1999), no. 3, 247–258.

[Si] B. Siebert, *Gromov-Witten invariants for general symplectic manifolds*, preprint, 1996.

[ST] B. Siebert and G. Tian, *On the holomorphicity of genus two Lefschetz fibrations*, preprint, 2001.

[Ta] C. Taubes, *Seiberg Witten and Gromov invariants for symplectic 4-manifolds.* Edited by Richard Wentworth. First International Press Lecture Series, 2. IP, 2000.

[Th] R. Thomas, *A holomorphic Casson invariant for Calabi-Yau 3-folds and bundles on $K3$ fibrations*, J. Differential Geom., 54 (2000), no. 2, 367–438.

[ThY] R. Thomas and S.T. Yau, *Special Lagrangians, stable bundles and mean curvature flow*, preprint, 2001.

[Ti1] G. Tian, *On Calabi's conjecture for complex surfaces with positive first Chern class*, Inv. Math., 101 (1990), no. 1, 101–172.

[Ti2] G. Tian, *Kähler-Einstein metrics with positive scalar curvature*, Invent. Math., 130 (1997), 1–39.

[Ti3] G. Tian, *Kähler-Einstein metrics on algebraic manifolds*, Proceedings of the International Congress of Mathematicians, Vol. I (Kyoto, 1990), 587–598.

[Ti4] G. Tian, *On a set of polarized Kähler metrics on algebraic manifolds*, J. Diff. Geom., 32 (1990), 99–130.

[Ti5] G. Tian, *Gauge theory and calibrated geometry. I*, Ann. of Math., (2) 151 (2000), no. 1, 193–268.

[Ti6] G. Tian, *Symplectic isotopy in four dimension*, First International Congress of Chinese Mathematicians (Beijing, 1998), 143–147, AMS/IP Stud. Adv. Math., 20, Amer. Math. Soc., 2001.

[Ts] H. Tsuji, *Existence and degeneration of Kähler-Einstein metrics on minimal algebraic varieties of general type*, Math. Ann., 281 (1988), no. 1, 123–133.

[TT] T. Tao and G, Tian, *A singularity removal theorem for Yang-Mills fields in higher dimensions*, preprint, 2001.

[TY1] G. Tian and S.T. Yau, *Existence of Kähler-Einstein Metrics on complete Kähler manifolds and their applications to algebraic geometry*, Mathematical aspects of string theory (San Diego, Calif., 1986), 574–628, Adv. Ser. Math. Phys., 1, World Sci. Publishing, Singapore, 1987.

[TY2] G. Tian and S.T. Yau, *Complete Khler manifolds with zero Ricci curvature. I*, J. Amer. Math. Soc., 3 (1990), no. 3, 579–609.

[TY3] G. Tian and S.T. Yau, *Complete Khler manifolds with zero Ricci curvature. II*, Invent. Math., 106 (1991), no. 1, 27–60.

[TYa] G. Tian and B.Z. Yang, *Compactification of the moduli spaces of vortices and coupled vortices*, preprint, 2002.

[TZ] G. Tian and X.H. Zhu, *Uniqueness of Khler-Ricci solitons*, Acta Math., 184 (2000), no. 2, 271–305.

[Uh1] K.K. Uhlenbeck, *Connections with L^p Bounds on Curvature*, Comm. in Math. Phys., 83 (1982), 31–42.

[Uh2] K.K. Uhlenbeck, *Removable Singularities in Yang-Mills Fields*, Comm. in Math. Phys., 83 (1982), 11–29.

[UY] K. Uhlenbeck and S.T. Yau, *On the existence of Hermitian-Yang-Mills connections in stable vector bundles*, Comm. Pure Appl. Math., 39 (1986), no. S, suppl., S257–S293.

[WaM] M.T. Wang, *Mean curvature flow of surfaces in Einstein four-manifolds*, J. Differential Geom., 57 (2001), no. 2, 301–338.

[Wang] McKenzie Y. Wang, *Einstein metrics from symmetry and bundle constructions*, Surveys in differential geometry: Essays on Einstein manifolds, 287–325, Surv. Differ. Geom., VI, Int. Press, Boston, MA, 1999.

[Ya] S.T. Yau, *On the Ricci curvature of a compact Kähler manifold and the complex Monge-Ampére equation, I^**, Comm. Pure Appl. Math., 31 (1978), 339–441.

[Zel] S. Zelditch, *Szegö kernels and a theorem of Tian*, Internat. Math. Res. Notices, 1998, no. 6, 317–331.

ICM 2002 · Vol. I · 495–504

Singularities in String Theory

E. Witten*

Abstract

String theory is a quantum theory that reproduces the results of General Relativity at long distances but is completely different at short distances. Mathematically, string theory is based on a very new — and little understood — framework for geometry that reduces to ordinary differential geometry when the curvature is asymptotically small. In the 1990's, many interesting results were obtained about the behavior of string theory in spacetimes that develop singularities. In many cases, the physics at the singularity is governed by an effective Lagrangian constructed using an interesting bit of classical geometry such as the association of A-D-E groups with certain hypersurface singularities or the ADHM construction of instantons. In other examples, the physics at the singularity cannot be described in classical terms but involves a non-Gaussian conformal field theory.

2000 Mathematics Subject Classification: 51P05, 81T30.
Keywords and Phrases: String theory, Quantum field theory, Mathematical physics.

1. Introduction

The classical Einstein equations

$$R_{IJ} = 0,$$

where R is the Ricci tensor of a metric g on spacetime, are scale-invariant. In other words, they are invariant under the scaling of the metric $g \to tg$, with t a real number; the Ricci tensor is invariant under this scaling. A quantum theory of gravity, however, cannot have this symmetry, since quantum theory depends on the *action*, not just the field equations, and the Einstein-Hilbert action

$$I = \frac{1}{16\pi G}\int d^n x\sqrt{g}R$$

* Institute For Advanced Study, Princeton NJ 08540, USA. E-mail: witten@sns.ias.edu

(G is Newton's constant and n is the dimension of spacetime) is not invariant under scaling. In fact, under $g \to tg$, the action scales as $I \to t^{n/2-1}I$, and so is not scale-invariant in any dimension above two. (Two dimensions — a case that figures in string theory — is completely special as I is a topological invariant, the Gauss-Bonnet integral.)

Generally speaking, the classical limit of quantum mechanics arises by making a stationary phase approximation to a function space integral. That integral is very roughly of the form

$$\int D\Phi \ \exp(iI/\hbar),$$

where the integral runs over all fields Φ. For example, for General Relativity, Φ would be a metric tensor on spacetime, perhaps together with other fields, and I would be the Einstein-Hilbert action defined in the last paragraph, together with a suitable action for the other fields, if any. Here $\hbar$ is Planck's constant, and the stationary phase approximation to the function space integral is valid when the appropriate $I/\hbar$ is large. Note that in classical physics, I is not a dimensionless number but has units of "action" or energy times time; it is only with the passage to quantum mechanics that there is a natural constant of action, namely $\hbar$, and it makes sense to say that in a given physical situation the action is large or small.

As an example of this criterion, consider black holes. A classical black hole can, because of the scale invariance, have any possible mass M or radius r; in four dimensions, for example, the mass and radius are related (for neutral, unrotating black holes) by $M = rc^2/G$, where c is the speed of light. The value of $I/\hbar$ integrated over the relevant time, which is the time for light to cross the black hole, is

$$\frac{I}{\hbar} = Mc^2 \cdot \frac{r}{c} \cdot \frac{1}{\hbar} = \frac{GM^2}{\hbar c} = \frac{r^2c^3}{G\hbar}.$$

The classical description of black holes is valid if this expression is large, or in other words if $M >> 10^{-5}$ gm or $r >> 10^{-33}$ cm.

Known astrophysical black holes have masses comparable to that of the sun (about 10^{33} gm) and above, so unless we get lucky with mini-black holes left over from the Big Bang, we are not going to be able to observe what happens for black holes so light and small that the classical description fails. But curiosity compels us to ask how to describe a small black hole, or what happens near the Big Bang, where classical General Relativity breaks down for similar reasons to what I have just described for Black Holes.

Here we run into a problem. One can read a textbook recipe for quantization in Dirac's old book or in more modern texts on quantum field theory. But these recipes, applied to the Einstein-Hilbert theory, do not work. Because of the highly nonlinear nature of Riemannian geometry, these methods fail to give a consistent and meaningful result.

This problem is very hard to convey to a mathematical audience because the whole question is about quantum field theory, which is not in clear focus as a

mathematical subject. The foundation for our modern understanding of elementary particles and forces is the success in quantizing theories such as Yang-Mills theory, whose action is

$$I = \frac{1}{4e^2} \int \mathrm{Tr}\, F \wedge *F,$$

where F is the curvature of a connection, e is a real constant, known as the gauge coupling constant, and Tr is an invariant quadratic form on the Lie algebra of the gauge group. Other somewhat analogous theories such as the quantum theory of maps from a Riemann surface to a fixed Riemannian manifold have also been extensively explored by physicists, with applications to both string theory and condensed matter physics. Apart from their central role in physics, quantum gauge theory and its cousins are the basis for the application of physical ideas to a whole range of mathematical problems, from the Jones polynomial to Donaldson theory and mirror symmetry.

But the understanding of quantum field theory that physicists have gained, though convincing and sufficient to make many computations possible, is hard to formulate rigorously. This makes it difficult for mathematicians to understand the questions of physicists, much less the partial results and approaches to a solution. In "constructive field theory," some of the standard physical claims about quantum field theory have been put on a rigorous basis, but there is still a long way to go to effectively bridge the gap.

For physicists, quantum gravity is an important problem not only because we would like to understand black holes, the Big Bang, and the quantum nature of spacetime — but also because reconciling General Relativity and quantum mechanics is necessary if we are to unify the forces of nature. We cannot achieve a unified understanding of nature if gravity is understood one way and the subatomic forces are understood by a different and incompatible theory. Moreover, the difficulty in reconciling these theories is probably our best clue about understanding physics at a much deeper level than we understand now. So what has been accomplished toward reconciling General Relativity and quantum mechanics?

2. String theory

Not much has been learned by direct assault. But roughly thirty years ago, in trying to solve another problem, physicists stumbled in "string theory" onto a very rich and surprising new framework for physics and geometry, which apparently does yield a theory of quantum gravity, though we do not understand it very well yet. String theory introduces in physics a new constant $\alpha' \sim (10^{-32}\,\mathrm{cm})^2$ (read "alpha-prime"), which is somewhat analogous to Planck's constant $\hbar \sim 10^{-27}\,\mathrm{erg\,sec}$, and modifies the concepts of physics in an equally far-reaching way.

If string theory is correct, then both $\hbar$ and α' are nonzero in nature. The deformation of classical physics in turning on nonzero $\hbar$ is comparatively familiar to most in this audience, at least at the level of nonrelativistic quantum mechanics (as opposed to quantum field theory): classical concepts such as the position

and velocity of a particle become "fuzzy" in the transition to quantum mechanics. Turning on $\alpha' \neq 0$ introduces an additional fuzziness in physics, roughly as a result of turning particles into strings. One aims to unify the forces by interpreting all of the different particles in nature as different vibrational states of one basic string.

String theorists spend the late 1980's and early 1990's largely studying the α' deformation. This work is hard to describe mathematically because it is all based on techniques of quantum field theory. Roughly it involves a new kind of geometry in which one is not allowed to talk about points or geodesics but one can talk about (quantum) minimal surfaces. This blurs all the classical concepts in geometry and makes possible nonclassical behavior. The new fuzziness has a characteristic scale $\sqrt{\alpha'} \sim 10^{-32}\,\mathrm{cm}$, and has many consequences. On much larger scales, just like the quantum uncertainty in gravity, the stringy fuzziness is unimportant. But it is important if one looks closely and can lead to nonclassical behavior, such as mirror symmetry.

A commonly encountered framework for nonclassical behavior in string theory is the following. Consider a family of classical solutions of string theory depending on several parameters; call the parameter space $\mathcal{N}$. At a generic point in $\mathcal{N}$, the stringy effects are important and one cannot usefully describe the situation in terms of a classical spacetime. As one approaches a special point $P \in \mathcal{N}$ (or more generally some locus in $\mathcal{N}$ of positive codimension), the relevant length scales in spacetime become large, the string effects become unimportant, and a classical spacetime X emerges. (P is typically a cusp-like point in $\mathcal{N}$; that is, there is typically a natural metric on $\mathcal{N}$, and P is at infinite distance in this metric.) In that limit, the string equations reduce to the classical Einstein equations on X (or more precisely, their appropriate supersymmetric extension, about which more later). To run what I have said so far in reverse, starting with a classical spacetime X that is embedded in string theory, if the radius of X (and every relevant length scale) is large compared to $\sqrt{\alpha'}$, then classical geometry is a good approximation to the stringy situation. But by varying the parameters so that the "radius" of X is not large compared to $\sqrt{\alpha'}$, one can get a situation in which classical geometry is not a good approximation and must be replaced with stringy geometry. This situation is described by points in the interior of $\mathcal{N}$.

A more seriously nonclassical behavior arises if in addition to the point P, there are additional points $Q, R \in \mathcal{N}$ at which classical behavior again arises, but this time with different classical spacetimes Y, Z of different topology. In this case, by moving in $\mathcal{N}$ from P to Q, we can go smoothly from a situation in which classical geometry is a good approximation and the spacetime is X, to a situation in which classical geometry is again a good approximation but the spacetime is Y. For this process to occur smoothly even though the initial and final spacetimes have different topology, it inevitably happens that in interpolating from P to Q, one has to pass through a region in which classical geometry is not a good approximation.

The example of what I have just described that is most discussed mathematically is that in which X is a Calabi-Yau threefold, and, say, Y is mirror to X and

Z is another Calabi-Yau manifold that is birational to X or Y.

Because the characteristic length scale of stringy behavior, in the simplest way of matching string theory with the real world, is about 10^{-32} cm, way below the distance scale that we can probe experimentally, much of the structure of string theory, assuming it is right, is out of reach experimentally. Conceivably, we might one day be able to use string theory to calculate masses and interaction rates of the observed elementary particles, but this seems far off. It is also just possible that we might have the chance to observe one or another kind of massive string relic left over from the early universe. There is, however, another possibility that is much more likely for the immediate future; there is one important aspect of stringy geometry which may very well be accessible to experiments. This is "supersymmetry," roughly the notion that spacetime is more accurately understood as a supermanifold, with both odd and even coordinates, rather than as an ordinary manifold. The theory of supermanifolds is more accessible and better known mathematically than some of the other things that I have mentioned. Most of the known geometrical applications of quantum field theory involve supersymmetry in one way or another.

If supersymmetry is relevant in nature, the oscillations of known particles in the "odd" directions in spacetime would give new elementary particles that could be discovered in accelerators. There are hints that these new particles exist at energies very close to what has already been reached experimentally. To me the most striking hint of this comes from the measured values of the strong, weak, and electromagnetic coupling constants, which are in excellent agreement with a prediction based on supersymmetric grand unification. If these hints have been correctly interpreted, we are likely to discover supersymmetric particles at accelerators in this decade, probably at the Fermilab accelerator in Illinois or at the Large Hadron Collider, which is being built at the European laboratory CERN near Geneva.

It is interesting to contemplate the impact on mathematics if supersymmetry is really discovered experimentally. When General Relativity emerged as an improvement on Newton's theory of gravity, this gave a huge boost to the mathematical investigation of Riemannian geometry. Nonrelativistic quantum mechanics probably gave an equivalent boost to functional analysis. Quantum field theory is so multi-faceted that a simple summary of its mathematical influence is difficult; some aspects of quantum field theory have influenced mathematics considerably, but as I have explained, problems of rigor have kept the core ideas of this richest of physical theories inaccessible mathematically. Supersymmetry, I think, would fall somewhere in between. Its experimental discovery would greatly increase the interest of mathematicians in supermanifold theory, which is accessible mathematically, but the full impact on mathematics would be delayed because the real payoff of supersymmetry lies in the realm of quantum field theory and string theory.

3. Uniform breakdown of the "large, smooth" approximation

In recent years, the most significant development in string theory has been to understand some of the things that happen when both $\hbar$ and α' are important. One discovery is that the different models of string theory that we knew of in the 1980's are related like the different classical spacetimes X, Y, and Z that we discussed above. In an asymptotic expansion near $\hbar = 0$, they are different, but in a more complete description, the different string models arise as different semiclassical limits of one richer theory that has been dubbed M-theory. M-theory, though we do not really understand it yet, is thus the candidate for super-unification of the laws of nature.

Also, we have gained insight about what happens in many situations in which classical geometry breaks down. In the "large, smooth" limit of spacetime, in which all relevant length scales are large, classical geometry (enriched to include supersymmetry) is always a good approximation. But what happens when the "large, smooth" approximation breaks down?

Many interesting results were obtained in the 1990's about situations in which the "large, smooth" approximation breaks down everywhere at once. I will give a few examples. These examples involve the basic ten-dimensional models of string theory, such as Type IIA and Type IIB superstrings and the heterotic string, and also the eleven-dimensional M-theory. Let $\mathbf{S}^1(r)$ denote a circle of circumference $2\pi r$. Then our first example is the assertion for any ten-dimensional spin manifold X, Type IIA superstring theory on $X \times \mathbf{S}^1(r)$ is equivalent to Type IIB superstring theory on $X \times \mathbf{S}^1(\alpha'/r)$. If $r >> (\alpha')^{1/2}$, the description via Type IIA superstring theory is transparent as ordinary geometrical concepts are valid, while for small r the second description is better. Starting on the Type IIA side at large r, the "large, smooth" description breaks down for $r \to 0$ (as there are closed geodesics of length $2\pi r$ in $X \times \mathbf{S}^1(r)$), and the equivalence to Type IIB on $X \times \mathbf{S}^1(\alpha'/r)$ gives a description that is valid when the Type IIA description has failed.

My other examples will relate an eleven-dimensional description via M-theory to a ten-dimensional string theory. With X as before a ten-dimensional spin manifold and Y a seven-dimensional spin manifold, and letting $\mathrm{K3}(r)$ denote a K3 surface of radius r and $\mathbf{I}(r)$ a length segment of length r, and $\mathbf{T}^3$ a three-torus, we have the following relations: (i) M-theory on $X \times \mathbf{S}^1(r)$ is equivalent as $r \to 0$ to Type IIA superstring theory on X; (ii) M-theory on $Y \times \mathrm{K3}(r)$ is equivalent as $r \to 0$ to the heterotic string on $Y \times \mathbf{T}^3$; and M-theory on $X \times \mathbf{I}(r)$ is equivalent for $r \to 0$ to the $E_8 \times E_8$ heterotic string on X. In each of these examples, the "large, smooth" approximation is valid for large r (if X and Y are large enough) and breaks down for small r. In each example, the string coupling constant in the string theory description vanishes for $r \to 0$, so that the string theory description is useful in that limit — an asymptotic expansion valid for small r can be explicitly worked out, giving a detailed answer to the question of what happens when the "large, smooth"

approximation fails. Note that these examples involve highly nonclassical behavior, with change in the topology and even the dimension of spacetime — for example, a four-dimensional K3 surface at large r is replaced by a three-dimensional torus when r becomes small.

Relations of this type are "quantum" analogs (involving both $\hbar$ and α') of mirror symmetry (which from this standpoint involves only α') and have led to the understanding that the different string models are different limits of the same things. Many other examples have been worked out in which the "large, smooth" approximation breaks down everywhere at once. I want to focus in the remaining time today, however, on another type of situation. This is the case in which, as some parameter is varied, the "large, smooth" approximation remains valid generically, but breaks down along some locus of codimension $d > 0$ where spacetime develops a conical singularity.

4. Behavior at conical singularities

The behavior of string theory when spacetime develops a conical singularity in positive codimension can be investigated by methods that exploit the fact that the "large, smooth" approximation remains generically valid, away from the singularity. One often can identify an interesting mathematical and physical phenomenon supported at the singularity. I will select examples in which both $\hbar$ and α' play an important role. There also are many instances of conical singularities that can be studied at $\hbar = 0$, such as the string theory orbifolds that have motivated one of the satellite meetings of ICM-2002. But we will focus on problems that involve both α' and $\hbar$. I will give three examples; two involve known mathematical constructions that appear in a new situation, while in the third the key phenomenon is nonclassical — it can only be formulated quantum mechanically.

I. *M-Theory At An A-D-E Singularity:* The A-D-E singularities are codimension four singularities that look locally like $\mathbf{R}^4/\Gamma$, where Γ is a finite subgroup of $SU(2)$, acting on $\mathbf{R}^4 \cong \mathbf{C}^2$ and preserving the hyper-Kahler structure of $\mathbf{R}^4$. An extensive mathematical theory relates the A-D-E singularity to the A-D-E Dynkin diagram and many associated bits of geometry and algebra. However, the role of the A-D-E *group* in relation to the singularity is elusive. Like other singular spaces, the A-D-E singularity is usefully studied as a limit of smooth spaces carrying the appropriate structure. In this example, the singular space $\mathbf{R}^4/\Gamma$ has a hyper-Kahler resolution (due to Kronheimer) that contains exceptional divisors of area $A_1, \dots, A_r$ which appear as parameters in the metric (r is the rank of the relevant A-D-E group, and the intersection form of the divisors is minus the Cartan matrix of the group). In the context of M-theory or string theory, for $A_1, \dots, A_r$ large, the "large, smooth" approximation is valid and classical geometry can be applied. We want to know what happens as $A_1, \dots, A_r \to 0$, giving a singularity at the origin in $\mathbf{R}^4$. This is the basic A-D-E singularity in $\mathbf{R}^4$; in eleven-dimensional M-theory, we would usually be working on an eleven-dimensional spacetime X, and the singularity arises

on a codimension-four submanifold Q. The answer has turned out to be that in this limit, A-D-E gauge fields — and their supersymmetric extension — appear on Q. Assuming that the "large, smooth" approximation has failed only because of the A-D-E singularity, there is an effective description of the resulting physics that roughly speaking is governed by the action

$$I = \int_X d^{11}x\sqrt{g}\,(R + \ldots) + \int_Q \mathrm{Tr}\,(F \wedge *F + \ldots)\,,$$

where R is the Ricci scalar, F is the curvature of the A-D-E connection, and "..." refers to the supersymmetric extension. The supersymmetric extension of the gauge theory that is supported on Q turns out to automatically contain the variables needed to parametrize Kronheimer's hyper-Kahler resolution of the singularity.

II. *Gauge Theory Instantons:* My second example involves the instanton solutions of four-dimensional Yang-Mills theory. Like the Einstein equations, the equations for Yang-Mills instantons, which read $F = -*F$, where F is the curvature of a Yang-Mills connection on $\mathbf{R}^4$, are scale invariant. (In fact, they have the much stronger property of conformal invariance.) Instantons therefore come in all sizes. One can scale the size of an instanton all the way down to zero, giving, in the limit, a singular, point-like instanton. So (even on a compact four-manifold) instanton moduli space is non-compact. This raises the question, "What happens when an instanton becomes small?"

The answer to this question depends on what one is trying to do. I will describe three possible answers. *(i)* Instantons were introduced in quantum field theory in the mid-1970's. Traditionally, physicists were interested in certain integrals over instanton moduli space. From this point of view, the meaning of the noncompactness of instanton moduli space is clear: one should make sure that the integrals of interest converge, and one should be careful when integrating by parts. *(ii)* In Donaldson theory, one is interested in intersection theory on instanton moduli space; "instanton bubbling" — the shrinking of an instanton to a point — is the main source of technical difficulty. One deals with it by a variety of technical means such as considering cycles in moduli space whose intersections avoid the bubbling region. *(iii)* In string theory, one expects the classical instanton equation $F = -*F$ to be a good approximation for large instantons, this being an example of the validity of classical concepts in the "large, smooth" region. But one expects this description to break down as the instanton shrinks. The question here is to find a description that is valid for small instantons.

The instanton problem can be embedded in string theory in different ways, so there are several answers. I will give the answer in one case — Type I superstring theory or the $SO(32)$ heterotic string. (At the very end of this talk, I briefly point out a second case.)

Before going on, I should mention one surprising part of the mathematical theory of instantons. This is the ADHM construction (due to Atiyah, Drinfeld, Hitchin, and Manin) of instantons in $\mathbf{R}^4$. To describe a k-instanton solution of

$SU(N)$ gauge theory on $\mathbf{R}^4$, the ADHM construction employs an auxiliary $U(k)$ group. (For example, the instanton moduli space is constructed as a hyper-Kahler quotient of a linear space divided by $U(k)$.) The interpretation of this group is somewhat mysterious in classical geometry, just at the role of the A-D-E group in relation to the A-D-E singularity is somewhat mysterious classically.

Instanton bubbling occurs at a point in $\mathbf{R}^4$, so in gauge theory in any dimension, it occurs on a submanifold of codimension four. In ten-dimensional Type I superstring theory on a ten-manifold X, the small instanton thus appears on a codimension four submanifold Q. The answer to the small instanton problem turns out to be that the $U(k)$ group of the ADHM construction appears as a gauge group in the physics on Q. The effective action that governs this situation turns out to be schematically

$$I = \int_X d^{11}x\sqrt{g}\,(R + \ldots) + \int_Q \mathrm{Tr}\,(F \wedge *F + \ldots)\,,$$

where in this case F is the curvature of a $U(k)$ connection, while "..." refers to additional terms required by supersymmetry plus the additional variables used in the ADHM construction to describe the moduli space as a hyper-Kahler quotient.

So once again, an interesting and surprising bit of classical mathematics becomes important near the singularity. I move on now, however, to an example in which the key phenomenon cannot be described in classical terms.

III. *Type IIB At An A-D-E Singularity:* Here we consider again the A-D-E singularity, but now in Type IIB superstring theory rather than in M-theory. The answer turns out to be completely different: we do not get a description with new classical degrees of freedom; instead a "non-trivial" or "non-Gaussian" conformal field theory is supported on the locus Q of the A-D-E singularity. The assertion that this theory is "non-trivial" means that it exists as a conformally invariant quantum field theory, but cannot be conveniently described in terms of classical or Gaussian fields.

This particular nontrivial conformal field theory might be described as a "nonabelian gerbe theory"; it is related to two-forms in roughly the way that nonabelian gauge theory is related to one-forms. Classically, one-forms have a nonabelian generalization in gauge theory, but to find an analogous theory for two-forms one must apparently go to quantum theory. The existence and basic properties of this particular six-dimensional conformal field theory can be used to deduce Montonen-Olive duality of quantum Yang-Mills theory in four dimensions, and this in turn has implications for certain four-manifold invariants.

So this example again involves interesting mathematics, but to describe the result requires use of quantum concepts in a more intimate way. The same happens if we consider the small instanton problem in the $E_8 \times E_8$ heterotic string (rather than Type I or the $SO(32)$ heterotic string as considered above). There is no ADHM construction for E_8 instantons, so there is no candidate for an answer along the lines

sketched above for Type I; instead, a non-Gaussian conformal field theory appears on the small instanton locus Q.

A general orientation to the subject matter discussed in this lecture can be found in the second half of volume 2 of [1]. A few of the original research papers of relevance are [2] for the A-D-E singularities, [3] for the ordinary double point singularity in complex dimension three (we have not actually discussed this case in the present lecture, but it was important in the development of the ideas), and [4] for small instantons. In addition, I have discussed the small instanton problem from a different but related point of view in [5]. Some readers may also want to consult general expositions of quantum field theory, such as the recent textbook [6] for physicists, or the exposition aimed at mathematicians in [7]. Finally, a comparatively recent account of known rigorous results on quantum field theory can be found in [8].

References

[1] J. Polchinski, *String Theory*, Cambridge University Press.

[2] E. Witten, String Theory Dynamics In Various Dimensions, *Nuclear Physics*, B443 (1995), 85.

[3] A. Strominger, Massive Black Holes and Conifolds In String Theory, *Nuclear Physics*, B451 (1995), 96.

[4] E. Witten, Small Instantons In String Theory, *Nuclear Physics*, *Nuclear Physics*, B460 (1996), 541.

[5] E. Witten, Small Instantons In String Theory, in *Prospects In Mathematics*, ed. H. Rossi, American Mathematical Society (1999), 111.

[6] S. Weinberg, *The Quantum Theory Of Fields*, Cambridge University Press.

[7] P. Deligne et. al, eds., *Quantum Fields and Strings: A Course For Mathematicians*, American Mathematical Society (1999).

[8] V. Rivasseau, *From Perturbative To Constructive Renormalization*, Princeton University Press (1991).

Appendix A: Invited Forty-Five Minute Lectures at the Section Meetings

(not included in Volumes II and III)

ICM 2002 · Vol. I · 507–513

The Power Set Function*

Moti Gitik†

Abstract

We survey old and recent results on the problem of finding a complete set of rules describing the behavior of the power function, i.e. the function which takes a cardinal κ to the cardinality of its power 2^κ.

1. Introduction

One of the central topics of Set Theory since Cantor was the study of the power function . The basic problem is to determine all the possible values of 2^κ for a cardinal κ. Paul Cohen [1] proved the independence of the Continuum Hypothesis and invented the method of forcing. Shortly after, Easton [3] building on Cohen's results showed the function $\kappa \longrightarrow 2^\kappa$, for regular κ, can behave in any prescribed way consistent with König's Theorem. This reduces the study to singular cardinals. It turned out that the situation with powers of singular cardinals is much more involved. Thus, for example, a remarkable theorem of Silver [22] states that a singular cardinal of uncountable cofinality cannot be first to violate GCH. The Singular Cardinal Problem is the problem of finding a complete set of rules describing the behavior of the power function on singular cardinals. There are three main tools for dealing with the problem: pcf-theory, inner models theory and forcing involving large cardinals.

2. Classical results and basic definitions

In 1938 Gödel proved the consistency of the Axiom of Choice (AC) and the Generalized Continuum Hypothesis (GCH) with the rest axioms of set theory. In 1963 Cohen proved the independence of AC and GCH. He showed, in particular, that $2^{\aleph_0}$ can be arbitrary large. Shortly after Solovay proved that $2^{\aleph_0}$ can take any value λ with $cf(\lambda) > \aleph_0$. The *cofinality* of a limit ordinal α ($cf(\alpha)$) is the least ordinal $\beta \leq \alpha$ so that there is a function $f : \beta \longrightarrow \alpha$ with $rng(f)$ unbounded in

*Partially supported by the Isreal Science Foundation.

†School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel. E-mail: gitik@post.tau.ac.il

α. A cardinal κ is called a *regular* if $\kappa = cf(\kappa)$. Otherwise a cardinal is called a *singular* cardinal. Thus, for example, $\aleph_8$ is regular and $\aleph_\omega$ is singular of cofinality ω.

By a result of Easton [3], if we restrict ourselves to regular cardinals, then every class function $F : Regulars \longrightarrow Cardinals\ satisfying$

(a) $\kappa \leq \lambda$ *implies* $F(\kappa) \leq F(\lambda)$

(b) $cf(F(\kappa) > \kappa$ (König's Theorem)

can be realized as a power function in a generic extension.

From this point we restrict ourselves to singular cardinals.

3. Restrictions on the power of singular cardinals

The Singular Cardinal Problem (SCP) is the problem of finding a complete set of rules describing the behavior of the power function on singular cardinals. For singular cardinals there are more limitations. Thus

(c) (Bukovsky - Hechler) If κ is a singular and there is $\gamma_0 < \kappa$ such that $2^\gamma = 2^{\gamma_0}$ for every $\gamma, \gamma \leq \gamma < \kappa$, then $2^\kappa = 2^{\gamma_0}$.

(d) (Silver) If κ is a singular strong limit cardinal of uncountable cofinality and $2^\kappa > \kappa^+$ then $\{\alpha < \kappa | 2^\alpha > \alpha^+\}$ contains a closed unbounded subset of κ.

A set $C \subset \kappa$ is called a *closed unbounded* subset of κ iff

(1) $\forall \alpha < \kappa \exists \beta \in C(\beta > \alpha)$ (*unbounded*)

(2) $\forall \alpha < \kappa(C \cap \alpha \neq \phi \Rightarrow sup(C \cap \alpha) \in C)$ (*closed*).

Subsets of κ containing a closed unbounded set form a filter over κ which is κ complete. A positive for this filter sets are called *stationry*.

(e) (Galvin - Hajnal, Shelah) If $\aleph_\delta$ is strong limit and $\delta < \aleph_\delta$ then $2^{\aleph_\delta} < \aleph_{2^{|\delta|^+}}$

(f) (Shelah) It is possible to replace $2^{|\delta|^+}$ in (e) by $|\delta|^{+4}$.

(g) (Shelah) Let $\aleph_\delta$ be the ω_1 -th fixed point of the $\aleph$ - function. If it is a strong limit , then $2^{\aleph_\delta} < min((2^{\omega_1})^+$ -fixed point, ω_4 -th fixed point).

A cardinal κ is called a *fixed point* of the $\aleph$ function if $\kappa = \aleph_\kappa$.

It is unknown if 4 in (f) and in (g) can be reduced or just replaced by 1. One of the major questions in Cardinal Arithmetic asks if $2^{\aleph_\omega}$ can be bigger than $\aleph_{\omega_1}$ provided it is a strong limit. We refer to the books by Jech [12] and by Shelah [23] for the proofs of the above results.

4. Inner models and large cardinals

There are other restrictions which depend on large cardinals. Thus the celebrated Covering Theorem of Jensen [2] implies that for every singular strong limit cardinal κ $2^\kappa = \kappa^+$, provided the universe is close to Gödel's model L (precisely, if $o\#$ does not exist, or, equivalently , there is no elementary embedding from L into L). On the other hand, using large cardinals (initially supercompact cardinals were used [14]) it is possible to have the following.

(Prikry-Silver, see [12]):

κ is a strong limit of cofinality ω and $2^\kappa > \kappa^+$.

(Magidor [15],[16],[17]):

(1) the same with κ of any uncountable cofinality.

(2) the same with $\kappa = \aleph_\omega$.

So, the answer to SCP may depend on presence of particular large cardinals. Hence, it is reasonable to study the possibilities for the power function level by level according to existence of particular large cardinals. There are generalizations of the Gödel model L which may include bigger and bigger large cardinals, have nice combinatorial properties, satisfy GCH and are invariant under set forcing extensions. This models are called *Core Models*. We refer to the book by Zeman [25] for a recent account on this fundamental results.

The Singular Cardinals Problem can now be reformulated as follows:

Given a core model K with certain large cardinals. Which functions can be realized in extensions of K as power set functions , i.e. let $F : Ord \longrightarrow Ord$ be a class function in K, is there an extension (generic) of K satisfying $2^{\aleph_\alpha} = \aleph_{F(\alpha)}$ for all ordinals α?

We will need few definitions.

An uncountable cardinal κ is called a *measurable* cardinal iff there is $\mu : P(\kappa) \longrightarrow \{0,1\}$ such that

(1) $\forall \alpha < \kappa \ \mu(\{\alpha\}) = 0$.

(2) $\mu(\kappa) = 1$.

(3) $A \subseteq B \Longrightarrow \mu(A) \leq \mu(B)$.

(4) $\forall \delta < \kappa \ \forall \ \{A_\nu | \nu < \delta\}$ subsets of κ with $\mu(A_\nu) = 0 \ \mu(\cup \ A_\nu | \nu < \delta\}) = 0$.

If κ is a measurable, then it is possible always to find μ with an additional property called *normality*:

(5) If $\mu(A) = 1$ and $f : A \longrightarrow \kappa, f(\alpha) < \alpha$ then there is a subset of A of measure one on which f is constant. Further by measure we shall mean a normal measure, i.e. one satisfying (1)–(5). A cardinal κ has the *Mitchell order* $\geq 1 (o(\kappa) \geq 1)$ iff κ is a measurable. A cardinal κ has the *Mitchell order* ≥ 2 $(o(\kappa) \geq 2)$ iff there is a measure over κ concentrating on measurable cardinals, i.e. $\mu(\{\alpha < \kappa | o(\alpha) \geq 1\}) = 1$.

In a similar fashion we can continue further , but up to κ^{++} only. Just the total number of ultrafilters over κ under GCH is κ^{++}. In order to continue above this point , directed systems of ultrafilters called extenders are used. This way we can reach κ with $o(\kappa) = Ord$. Such κ is called a *strong cardinal*. Core models are well developed to the level of strong cardinal and much further. Almost all known consistency results on the Singular Cardinals Problem require large cardinals below the level of a strong cardinal.

5. Finite gaps

By results of Jensen [2], Dodd- Jensen [13], Mitchell [20], Shelah [23] and Gitik [5] nothing interesting in sense of SCP happens bellow the level of $o(\kappa) = \kappa^{++}$. If there is $n < \omega$ such that for every $\alpha, o(\alpha) \leq \alpha^{+n}$, then we have the following additional restrictions:

(1) (Gitik-Mitchell [10]) If κ is a singular strong limit and $2^\kappa = \kappa^{+m}$ for some $m > 1$, then, in $K, o(\kappa) \geq \kappa^{+m}$. In particular, $m \leq n$.

(2) If κ is a singular cardinal of uncountable cofinality and for some $m, 1 \leq m < \omega$ $\{\alpha < \kappa | 2^\alpha = \alpha^{+m}\}$ is stionry, then $\{\alpha < \kappa | 2^\alpha = \alpha^{+m}\}$ contains a closed unbounded subset of κ.

By results of Merimovich [18] it looks like this are the only restrictions.

6. Uncountable cofinality case

Assume only that there is no inner model with a strong cardinal. Then we have the following restrictions:

(1) If κ is a singular strong limit cardinal of uncountable cofinality δ and $2^\kappa \geq \lambda > \kappa^+$, where λ is not the successor of a cardinal of cofinality less than κ, then $o(\kappa) \geq \lambda + \delta$, if $\delta > \omega_1$ or $o(\kappa) \geq \lambda$, if $\delta = \omega_1$.

(2) Let κ be a singular strong limit cardinal of uncountable cofinality δ and let $\tau < \delta$. If $A = \{\alpha < \kappa | cf\alpha > \omega, 2^\alpha = \alpha^{+\tau}\}$ is stionry, then A contains a closed unbounded subset of κ.

(3) If $\delta < \aleph_\delta, \aleph_\delta$ strong limit then $2^{\aleph_\delta} < \aleph_{|\delta|^+}$

(This was improved recently by R.Schindler [11] to many Woodin cardinals).

(4) Let $\aleph_\delta$ be the ω_1 -th fixed point of the $\aleph$-function. If it is a strong limit cardinal then $2^{\aleph_\delta} < \omega_2$ -th fixed point.

(5) If a is an uncountable set of regular cardinals with $min(a) > 2^{|a|^+ + \aleph_2}$, then $|pcf(a)| = |a|$, where $pcf(a) = \{cf(\Pi a/D) | D$ is an ultrafilter over $a\}$.

It is a major problem of Cardinal Arithmetic if it is possible to have a set of regular cardinals a with $min(a) > |a|$ such that $|pcf(a)| > |a|$. The results above were proved in Gitik-Mitchell [10], and in [7]. It is unknown if there is no further restrictions in this case (i.e. singulars of uncountable cofinality under the assumption that there is no inner model with a strong cardinal). Some local cases were checked by Segal [21] and Merimovich [19].

7. Countable cofinality case

In this section we revue some more recent results dealing with countable cofinality. First suppose that

$(\forall n < \omega \exists \alpha\ o(\alpha) = \alpha^{+n})$, but $\neg(\exists \alpha\ o(\alpha) = \alpha^{+\omega})$.

Then the following holds: Let κ be a cardinal of countable cofinality such that for every $n < \omega$ $\{\alpha < \kappa | o(\alpha) \geq \alpha^{+n}\}$ is unbounded in κ. Then for every $\lambda \geq \kappa^+$ there is a cardinal preserving generic extension satisfying "κ is a strong limit and $2^\kappa \geq \lambda$ ". So the gap between a singular and its power can already be arbitrary large . But by [7]:

If $2^\kappa \geq \kappa^{+\delta}$ for $\delta \geq \omega_1$, then GCH cannot hold below κ.

(Actually , GCH can hold if the gap is at most countable [8].)

We do not know if "pcf (a)uncountable for a countable a" is stronger than the assumption above. If we require also GCH below, then it is.

Once one likes to have uncountable gaps between a singular cardinal and its power together with GCH below , then the following results provide this and are sharp. The proofs are spread through papers [8], [9], [10].

Suppose that $\kappa > \delta \geq \aleph_0, \delta$ is a cardinal, $2^\kappa \geq \kappa^{+\delta}, cf\kappa = \aleph_0$ and GCH below κ. Then

(i) $cf\delta = \aleph_0$ implies (that in the core model) for every $\tau < \delta$ $\{\alpha < \kappa | o(\alpha) \geq \alpha^{+\tau}\}$ is unbounded in κ.

(ii) $cf\delta > \aleph_0$ implies (in the core model) $o(\kappa) \geq \kappa^{+\delta+1} + 1$ or $\{\alpha < \kappa | o(\alpha) \geq \alpha^{+\delta+1}\}$ is unbounded in κ.

Finally let us consider the following large cardinal: κ is singular of cofinality ω and for every $\tau < \kappa$ $\{\alpha < \kappa | o(\alpha) \geq \alpha^{+\tau}\}$ is unbounded in κ.

Under this assumption it is possible to blow up the power of κ arbitrary high preserving GCH below κ. Also, it is possible to turn κ into the first fixed point of the $\aleph$ function, see [6]. This answers Question (γ) from the Shelah's book on cardinal arithmetic [23]. What are the possibilities for the power function under the assumption above? First in order to be able to deal with cardinals above κ, let us replace it by a global one:

For every τ there is α $o(\alpha) \geq \alpha^{+\tau}$.

We do not know the status of "*pcf* of a countable set uncountable", but other limitations like

(1) $\aleph_\omega$ strong limit implies $2^{\aleph_\omega} < \aleph_{\omega_1}$

(2) If κ is a singular of uncountable cofinality then either $\{\alpha < \kappa | 2^\alpha \geq \alpha^+\}$ *or* $\{\alpha < \kappa | 2^\alpha > \alpha^+\}$ contains a closed unbounded subset of κ

are true below strong cardinal.

By recent result [11] the negation of the second assumption implies initially unrelated statement - Projective Determinacy. We refer to the books by A. Kanamori [14] and H. Woodin [24] on this subject. We conjecture that there is no other limitations, i.e. (1) with $\aleph_\omega$ replaced by $\aleph_\delta$ for $\delta < \aleph_\delta$, (2) and the classical ones.

8. One idea

Let us conclude with a sketch of one basic idea which is crucial for the forcing constructions in the countable cofinality case. Let U be a κ complete nontrivial ultrafilter over κ (say, in K). A sequence $\langle \delta_n | n < \omega \rangle$ is called a *Prikry sequence* for U iff for each $A \in U$ $\exists n_0 \forall n \geq n_0$ $\delta_n \in A$. Suppose now that κ is a strong limit singular cardinal of cofinality ω and $2^\kappa = \kappa^{++}$. Then, usually (by [5], [10]), we will have a sequence $\langle U_\alpha | \alpha < \kappa^{++} \rangle$ of ultrafilters in K and a sequence $\langle \delta_{\alpha,n} | \alpha < \kappa^{++}, n < \omega \rangle$ so that

(1) $\alpha < \beta \Longrightarrow \exists n_0 \forall n \geq n_0 \delta_{\alpha,n} < \delta_{\beta,n}$,

(2) $\langle \delta_{\alpha,n} | n < \omega \rangle$ is a Prikry sequence for U_α.

Ultrafilters U_α are different here. So each sequence $\langle \delta_{\alpha,n} | n < \omega \rangle$ relates to unique ultrafilter from the list. But once κ^{++} is replaced by κ^{+++}, the corresponding sequence of ultrafilters $\langle U_\alpha | \alpha < \kappa^{+++} \rangle$ will have different α and β, $\kappa^{++} < \alpha < \beta < \kappa^{+++}$ with $U_\alpha = U_\beta$. Then a certain Prikry sequence $\langle \delta_n | n < \omega \rangle$ may pretend to correspond to both U_α and U_β. In order to decide, we will need a

Prikry sequence for some U_γ with $\gamma < \kappa^{++}$ (more precisely, if f_β is the canonical one to one correspondence in K between κ^{++} and β then $f_\beta(\gamma) = \alpha$). Dealing with κ^{+4} we will need go down twice, first to κ^{+3} and after that to κ^{++}. In general, for $n, 3 \leq n < \omega$, $n-2$-many times. Certainly, it is impossible to go down infinitely many times, but instead we replace the fixed κ by an increasing sequence $\langle \kappa_n | n < \omega \rangle$ with each κ_n carrying κ_n^{+n+3} many ultrafilters. Now it turns out to be possible to add ω - sequences with no assignment to ultrafilters. Just the number of steps needed to produce the assignment is ω which is not enough for sequences of the length ω.

References

[1] P.Cohen, The independence of the continuum hypothesis, Proc. Nat. Acad. Sci. U.S.A. 50 (1963), 1143–1148; 51 (1964), 105–110.

[2] K.Devlin, Constructibility, Springer-Verlag, 1984.

[3] W. Easton, Powers of regular cardinals, Ann. Math. Logic 1 (1970), 139–178.

[4] F. Galvin and A. Hajnal , Inequalities for cardinal powers, Ann. of Math. 101 (1975), 491–498.

[5] M.Gitik, The strength of the failure of the Singular Cardinals Hypothesis, Ann. of Pure and App. Logic 51(3) (1991), 215–240.

[6] M.Gitik, No bound for the first fixed point, submitted.

[7] M.Gitik, On gaps under GCH type assumptions, to appear in Annals of Pure and Appl. Logic, math.LO/9908118.

[8] M.Gitik, Blowing up power of singular cardinal- wider gaps, Annals of Pure and Appl. Logic, 116 (2002) 1–38.

[9] M.Gitik and M.Magidor, The singular cardinals problem revisited, in: H.Judah, W.Just and W.H. Woodin, eds., Set Theory of the Continuum (Springer, Berlin, 1992), 243–316.

[10] M.Gitik and W.Mitchell, Indiscernible sequences for extenders, and the singular cardinal hypothesis, Annals of Pure and Appl. Logic 82 (1996), 273–316.

[11] M.Gitik, S. Shelah and R. Schindler, Pcf theory and Woodin cardinals, to appear.

[12] T.Jech, Set Theory, Springer-Verlag, 1997.

[13] A.Dodd and R.Jensen, The Core Model, Annals Math. Logic 20 (1981), no.1, 43–75.

[14] A.Kanamori, The Higher Infinite, Springer-Verlag, 1994.

[15] M. Magidor, Changing cofinality of cardinals, Fundamenta. Math. 99 (1978), 61–71.

[16] M. Magidor, On the singular cardinal problem 1, Isr. J. of Math. 28 (1977), 1–31.

[17] M. Magidor, On the singular cardinal problem 2, Ann. of Math.,106 (1977)517–547.

[18] C. Merimovich, A power function with a fixed finite gap, to appear in J. of Symbolic Logic.

[19] C. Merimovich, Extender based Radin forcing, to appear in Trans. AMS.

[20] W.Mitchell, Applications of the covering lemma for sequences of measures, Trans. AMS 299(1) (1987), 41–58.
[21] M. Segal, Master thesis, The Hebrew University, 1993.
[22] J. Silver, On the singular cardinals problem, Proc. ICM 1974, 265–268.
[23] S.Shelah, Cardinal Arithmetic, Oxford Logic Guides, vol. 29, Oxford University Press, Oxford, 1994.
[24] H.Woodin, Te Axiom of Determinacy, Forcing Axioms, and the Nonstationry Ideal, , de Gruyter Series in Logic and Its Applications, vol. 1, 1999.
[25] M.Zeman, Inner Models and Large Cardinals, de Gruyter Series in Logic and Its Applications, vol. 5, 2002.

ICM 2002 · Vol. I · 515–524

Beyond $\sum_1^2$ Absoluteness

W. Hugh Woodin*

Abstract

There have been many generalizations of Shoenfield's Theorem on the absoluteness of Σ_2^1 sentences between uncountable transitive models of ZFC. One of the strongest versions currently known deals with Σ_1^2 absoluteness conditioned on CH. For a variety of reasons, from the study of inner models and from simply combinatorial set theory, the question of whether conditional Σ_2^2 absoluteness is possible at all, and if so, what large cardinal assumptions are involved and what sentence(s) might play the role of CH, are fundamental questions. This article investigates the possiblities for Σ_2^2 absoluteness by extending the connections between determinacy hypotheses and absoluteness hypotheses.

2000 Mathematics Subject Classification: 03E45, 03E55, 03E10,04A10, 04A13.
Keywords and Phrases: Determinacy, Large cardinals, Forcing, Ω-logic.

1. Absoluteness and strong logics

There have been many generalizations of Shoenfield's Theorem on the absoluteness of Σ_2^1 sentences between uncountable transitive models of ZFC. Absoluteness theorems are meta-mathematically interesting since they identify levels of complexity where the technique of forcing cannot be used to establish independence.

A sentence, ϕ, is a Σ_1^2-*sentence* if for some Σ_1-formula, $\psi(x)$, ϕ is provably equivalent in ZFC, Zermelo Frankel set theory with the *Axiom of Choice*, to the assertion that $\psi[\mathbb{R}]$ holds. While this is not the standard definition, for the purposes of this article it is equivalent.

Theorem 1.1 *Suppose that ϕ is a Σ_1^2 sentence, there exists a proper class of measurable Woodin cardinals and that* CH *holds. Suppose $\mathbb{P}$ is a partial order and that $V^{\mathbb{P}} \vDash$ CH. Then $V \vDash \phi$ if and only if $V^{\mathbb{P}} \vDash \phi$* □

*Department of Mathematics, University of California, Berkeley, Berkeley CA 94720, USA. E-mail: woodin@math.berkeley.edu

This is Σ^2_1 generic absoluteness conditioned on CH. Because CH is itself a Σ^2_1 sentence, this *conditional* form of Σ^2_1 generic absoluteness is the best one can hope for. The meta-mathematical significance of this kind of absoluteness result is simply this. If a problem is expressible as a Σ^2_1 sentence, and there are many such examples from analysis, then it is likely settled by CH (augmented by modest large cardinal hypotheses). The technique of forcing cannot be used to demonstrate otherwise.

Absoluteness theorems can be naturally reformulated using *strong logics.* For generic absoluteness the relevant logic is Ω^*-logic.

Definition 1.2(Ω^*-logic) *Suppose that there exists a proper class of Woodin cardinals and that ϕ is a sentence. Then*

$$\mathrm{ZFC} \vdash_{\Omega^*} \phi$$

if for all ordinals α and for all partial orders $\mathbb{P}$ if $V^{\mathbb{P}}_\alpha \vDash \mathrm{ZFC}$, then $V^{\mathbb{P}}_\alpha \vDash \phi$. □

The theorem on Σ^2_1-absoluteness and CH can be reformulated as follows.

Theorem 1.3 *Suppose there exists a proper class of measurable Woodin cardinals. Then for each Σ^2_1 sentence ϕ, either* $\mathrm{ZFC+CH} \vdash_{\Omega^*} \phi$; *or* $\mathrm{ZFC+CH} \vdash_{\Omega^*} (\neg\phi)$.□

But there is another natural strong logic; Ω-logic, the definition of Ω-logic involves universally Baire sets of reals.

Definition 1.4 [1] *A set $A \subseteq \mathbb{R}^n$ is* universally Baire *if for any continuous function, $F : \Omega \to \mathbb{R}^n$, where Ω is a compact Hausdorff space, the preimage of A,*

$$\{p \in X \mid F(p) \in A\},$$

has the property of Baire in Ω; i. e. is open in Ω modulo a meager set. □

Every borel set $A \subseteq \mathbb{R}^n$ is universally Baire. More generally the universally Baire sets form a σ-algebra closed under preimages by borel functions

$$f : \mathbb{R}^n \to \mathbb{R}^m.$$

The universally Baire sets have the classical regularity properties of the borel sets, for example they are Lebesgue measurable and have the property of Baire. If there exists a proper class of Woodin cardinals then the universally Baire sets are closed under projection and every universally Baire set is determined.

Suppose that $A \subseteq \mathbb{R}$ in universally Baire and that $V[G]$ is a set generic extension of V. Then the set A has canonical interpretation as a set

$$A_G \subseteq \mathbb{R}^{V[G]}.$$

The set A_G is defined as

$$A_G = \cup\{\mathrm{range}(\pi_G) \mid \pi \in V, \mathrm{range}(\pi) = A\};$$

here π is a function, $\pi : \lambda^\omega \to \mathbb{R}$, which satisfies the uniform continuity requirement that for $f \neq g$;

$$|\pi(f) - \pi(g)| < 1/(n+1)$$

where $n < \omega$ is least such that $f(n) \neq g(n)$. If there exists a proper class of Woodin cardinals then

$$\langle H(\omega_1), A, \in\rangle \prec \langle H(\omega_1)^{V[G]}, A_G, \in\rangle.$$

Definition 1.5 *Suppose that $A \subseteq \mathbb{R}$ is universally Baire and that M is a transitive set such that $M \vDash$ ZFC. Then M is* A-closed *if for each partial order $\mathbb{P} \in M$, if $G \subset \mathbb{P}$ is V-generic then in $V[G]$: $A_G \cap M[G] \in M[G]$.* □

Definition 1.6(Ω logic) *Suppose that there exists a proper class of Woodin cardinals and that ϕ is a sentence. Then* ZFC $\vdash_\Omega \phi$ *if there exists a universally Baire set $A \subseteq \mathbb{R}$ such that if M is any countable transitive set such that $M \vDash$ ZFC and such that M is A-closed, then $M \vDash \phi$.* □

Both Ω-logic and Ω^*-logic are definable and generically invariant.

A natural question, given the theorem on generic absoluteness for Σ_1^2 is the following question:

Suppose there exists a proper class of measurable Woodin cardinals. Does it follow that for each Σ_1^2 sentence ϕ, either ZFC + CH $\vdash_\Omega \phi$; *or* ZFC + CH $\vdash_\Omega (\neg\phi)$?

The answer is yes if "iterable" models with measurable Woodin cardinals exist.

Ω Conjecture:

Suppose that there exists a proper class of Woodin cardinals and that ϕ is a Π_2 sentence. Then ZFC $\vdash_{\Omega^*} \phi$ *if and only if* ZFC $\vdash_\Omega \phi$.

It is immediate from the definitions and Theorem 1.1, that the Ω Conjecture settles the question above affimatively. But the consequences of the Ω-Conjecture are far more reaching. If the Ω Conjecture is true, then generic absoluteness is equivalent to absoluteness in Ω-logic and this in turn has significant metamathematical implications.

We fix some conventions. A formula, $\phi(x)$, is a Σ_2^2*-formula* if for some Σ_2-formula, $\psi(x)$, the formula $\phi(x)$ is provably equivalent in ZFC to the formula

$$\text{“}x \in H(c^+) \text{ and } \langle H(c^+), \in\rangle \vDash \psi[x]\text{”}.$$

Finally $\phi(x)$ is a $\Sigma_2^2(\mathcal{I}_{\text{NS}})$*-formula* if for some Σ_2-formula, $\psi(x)$, the formula $\phi(x)$ is provably equivalent in ZFC to the formula

$$\text{“}x \in H(c^+) \text{ and } \langle H(c^+), \mathcal{I}_{\text{NS}}, \in\rangle \vDash \psi[x]\text{”}.$$

where $\mathcal{I}_{\text{NS}}$ denotes the nonstationary ideal on ω_1.

There is a limit to the possible extent of absoluteness in Ω-logic. One version is given by the following theorem.

Theorem 1.7 *Suppose that there exist a proper class of Woodin cardinals, Ψ is a sentence and that for each $\Sigma_2^2(\mathcal{I}_{\text{NS}})$ sentence ϕ, either* ZFC + $\Psi \vdash_\Omega \phi$, *or* ZFC + $\Psi \vdash_\Omega (\neg\phi)$. *Then* ZFC + Ψ *is Ω-inconsistent.* □

In short:

$\Sigma^2_2(\mathcal{I}_{\rm NS})$ absoluteness is not possible in Ω-logic. If the Ω Conjecture holds then generic absoluteness for $\Sigma^2_2(\mathcal{I}_{\rm NS})$ sentences is not possible.

So for absoluteness in Ω-logic the most one can hope for is that there exist a sentence Ψ such that for each Σ^2_2 sentence ϕ, either $\mathrm{ZFC} + \Psi \vdash_\Omega \phi$, or $\mathrm{ZFC} + \Psi \vdash_\Omega (\neg\phi)$. In particular if the Ω Conjecture holds then Σ^2_2 generic absoluteness is the most one can hope for.

Suppose that Ψ is a sentence such that for each Σ^2_2 sentence ϕ, either $\mathrm{ZFC} + \Psi \vdash_\Omega \phi$, or $\mathrm{ZFC} + \Psi \vdash_\Omega (\neg\phi)$. Then $\mathrm{ZFC} + \Psi \vdash_\Omega 2^{\aleph_0} < 2^{\aleph_1}$. A natural conjecture is that in fact $\mathrm{ZFC} + \Psi \vdash_\Omega \mathrm{CH}$.

In any case from this point on we shall consider absoluteness just in the context of CH.

Generic absoluteness is closely related to determinacy. The statement of a theorem which illustrates one aspect of this requires the following definition.

Definition 1.8 *Suppose that there exists a proper class of Woodin cardinals. A set $A \subseteq \mathbb{R}$ is* Ω^*-recursive *if there exists a formula $\phi(x)$ such that:*

1. *$A = \{r \mid \mathrm{ZFC} \vdash_{\Omega^*} \phi[r]\}$;*
2. *For all partial orders, $\mathbb{P}$, if $G \subseteq \mathbb{P}$ is V-generic then for each $r \in \mathbb{R}^{V[G]}$, either*

$$V[G] \vDash \text{“}\mathrm{ZFC} \vdash_{\Omega^*} \phi[r]\text{”},$$

or $V[G] \vDash$ “$\mathrm{ZFC} \vdash_{\Omega^*} (\neg\phi)[r]$”. □

Theorem 1.9 *Suppose that there exists a proper class of Woodin cardinals. Suppose that $A \subseteq \mathbb{R}$ is Ω^*-recursive. Then A is determined.* □

On the other hand there are many examples where suitable determinacy assumptions imply generic absoluteness. Our main results deal with generalizations of these connections to Σ^2_1 and Σ^2_2 in the context of CH.

2. Absoluteness and determinacy

We fix a reasonable coding of elements of $H(\omega_1)$ by reals. This is simply a surjection

$$\pi : \mathrm{dom}(\pi) \to H(\omega_1)$$

where $\mathrm{dom}(\pi) \subseteq \mathbb{R}$. All we require of π is that $\pi \in L(\mathbb{R})$; the natural choice for π is definable in $H(\omega_1)$. For each set $X \subseteq H(\omega_1)$ let

$$X^* = \{x \in \mathbb{R} \mid \pi(x) \in X\}.$$

Suppose $X \subseteq \{0,1\}^{\omega_1}$. Associated to X is a game of length ω_1. The convention is that Player I plays first at limit stages. Strategies are functions:

$$\tau : \{0,1\}^{<\omega_1} \to \{0,1\}.$$

Suppose that $\Gamma \subseteq \mathcal{P}(\mathbb{R})$. Then X is Γ-*clopen* if there exist sets $Y \subset H(\omega_1)$ and $Z \subset H(\omega_1)$ such that

1. $Y \cap Z = \emptyset$,
2. for all $a \in \{0,1\}^{\omega_1}$ there exists $\alpha < \omega_1$ such that either $a|\alpha \in Y$ or $a|\alpha \in Z$,
3. X is the set of $a \in \{0,1\}^{\omega_1}$ such that there exists $\alpha < \omega_1$ such that $a|\alpha \in Y$ and such that $a|\beta \notin Z$ for all $\beta < \alpha$,
4. $Y^* \in \Gamma$ and $Z^* \in \Gamma$.

The first result on the determinacy of Γ-clopen sets is due to Itay Neeman. One version is the following.

Theorem 2.1 [2] *Suppose that there exists a Woodin cardinal which is a limit of Woodin cardinals. Suppose that $X \subseteq \{0,1\}^{\omega_1}$ and X is Π^1_1-clopen.*
Then X is determined. □

The proof of Neeman's theorem combined with techniques from the fine structure theory associated to AD^+ yields the following generalization which we shall need.

Theorem 2.2 *Suppose that there is a proper class of Woodin cardinals which are limits of Woodin cardinals. Let Γ^∞ be the set of all $A \subseteq \mathbb{R}$ such that A is universally Baire and suppose that $X \subseteq \{0,1\}^{\omega_1}$ is such that X is Γ^∞-clopen.*
Then X is determined. □

Suppose $X \subseteq \{0,1\}^{\omega_1}$ and that $\Gamma \subseteq \mathcal{P}(\mathbb{R})$. Then X is Γ-*open* if there exist sets $Y \subset H(\omega_1)$ such that

1. X is the set of $a \in \{0,1\}^{\omega_1}$ such that there exists $\alpha < \omega_1$ such that $a|\alpha \in Y$.
2. $Y^* \in \Gamma$.

John Steel has proved that under fairly general conditions, if $\Gamma \subseteq \mathcal{P}(\mathbb{R})$ is such that all Γ-open sets are determined then for each $X \subseteq \{0,1\}^{\omega_1}$, if X is Γ-open and if Player I wins the game given by X, then there is a winning strategy for Player I which is (suitably) definable from parameters in Γ; [4]. The following is a straightforward corollary:

Corollary 2.3 *Suppose that there exists a proper class of Woodin cardinals. Let Γ^∞ be the set of all $A \subseteq \mathbb{R}$ such that A is universally Baire and suppose that for each $A \in \Gamma^\infty$, $\mathrm{ZFC} \vdash_\Omega$ " All $\sum^1_1(A)$-open games are determined".*
Then for each $A \in \Gamma^\infty$, for each Σ^2_1-formula $\phi(x)$; either $\mathrm{ZFC} + \mathrm{CH} \vdash_\Omega \phi[A]$ or $\mathrm{ZFC} + \mathrm{CH} \vdash_\Omega (\neg\phi)[A]$. □

Using the theorem on the determinacy of Γ^∞-clopen games one obtains the converse.

Theorem 2.4 *Suppose that there exists a proper class of Woodin cardinals which are limits of Woodin cardinals. Let Γ^∞ be the set of all $A \subseteq \mathbb{R}$ such that A is universally Baire. Then the following are equivalent.*

(1) *For each $A \in \Gamma^\infty$, $\mathrm{ZFC} \vdash_\Omega$ " All $\sum^1_1(A)$-open games are determined".*
(2) *For each $A \in \Gamma^\infty$, for each Σ^2_1-formula $\phi(x)$, either $\mathrm{ZFC} + \mathrm{CH} \vdash_\Omega \phi[A]$ or $\mathrm{ZFC} + \mathrm{CH} \vdash_\Omega (\neg\phi)[A]$.* □

A set $A \subseteq \mathbb{N}$ is *Ω-recursive* if there exists a formula $\phi(x)$ such that for all $k \in \mathbb{N}$, either $\mathrm{ZFC} \vdash_\Omega \phi[k]$ or $\mathrm{ZFC} \vdash_\Omega (\neg\phi)[k]$; and such that

$$A = \{k \in \mathbb{N} \mid \mathrm{ZFC} \vdash_\Omega \phi[k]\}.$$

The question of whether there exists a sentence Ψ such that for each Σ^2_2 sentence ϕ, either $\mathrm{ZFC} + \mathrm{CH} + \Psi \vdash_\Omega \phi$, or $\mathrm{ZFC} + \mathrm{CH} + \Psi \vdash_\Omega (\neg\phi)$, and such that $\mathrm{ZFC} + \mathrm{CH} + \Psi$ is Ω-consistent; can be reformulated as:

> *Suppose there exists a proper class of Woodin cardinals and that* CH *holds. Let T be the set of all Σ^2_2-sentences, ϕ, such that*
>
> $$V \vDash \phi.$$
>
> *Can T be Ω-recursive?*

Theorem 2.5 *Suppose that there exists a proper class of inaccessible limits of Woodin cardinals. Let Γ^∞ be the set of all $A \subseteq \mathbb{R}$ such that A is universally Baire and suppose that all Γ^∞-open games are determined. Let $T_{\max}$ be the set of all Σ^2_2 sentences ϕ such that $\mathrm{ZFC} + \mathrm{CH} + \phi$ is Ω-consistent.*
Then $\mathrm{ZFC} + \mathrm{CH} + T_{\max}$ is Ω-consistent. □

The following conjecture can be proved from rather technical assumptions on the exsitence of an inner model theory for the large cardinal hypothesis: *κ is δ supercompact where $\delta > \kappa$ and δ is a Woodin cardinal.* The conjecture is:

> *Suppose that there exists a proper class of supercompact cardinals. Let $T_{\max}$ be the set of all Σ^2_2 sentences ϕ such that $\mathrm{ZFC} + \mathrm{CH} + \phi$ is Ω-consistent. Then $T_{\max}$ is Ω-recursive.*

While the plausibility of this conjecture is some evidence that Σ^2_2 absoluteness is possible, it does not connect Σ^2_2 absoluteness with any determinacy hypothesis.

From inner model theory considerations any such determinacy hypothesis must be beyond the reach of superstrong cardinals. In fact, Itay Neeman has defined a family of games whose (provable) determinacy is arguably beyond the reach of superstrong cardinals; [3].

3. Neeman games

For each formula $\phi(x_1, \ldots, x_n)$, let X_ϕ be the set of all $a \in \{0,1\}^{\omega_1}$ such that there exists a closed, unbounded set $C \subseteq \omega_1$ such that for all $\alpha_1 < \cdots < \alpha_n$ in C,

$$\langle H(\omega_1), a, \in\rangle \vDash \phi[\alpha_1, \ldots, \alpha_n].$$

The game given by X_ϕ is a *Neeman game.* Are Neeman games determined?

Surprisingly the consistency strength of the determinacy of all Neeman games is relatively weak.

Lemma 3.1 *Suppose that all Δ^1_3-clopen games are determined. Then there exists $A \subseteq \omega_1$ such that in $L[A]$ if $X \subseteq \{0,1\}^{\omega_1}$ is definable an ω-sequence of ordinals, then X is determined.* □

One can easily introduce additional predicates for sets of reals.

For each formula, $\phi(x_1, \ldots, x_n)$, and for each set $A \subseteq \mathbb{R}$ let $X_{(\phi,A)}$ be the set of all $a \in \{0,1\}^{\omega_1}$ such that there exists a closed, unbounded set $C \subseteq \omega_1$ such that for all $\alpha_1 < \cdots < \alpha_n$ in C, $\langle H(\omega_1), a, A, \in\rangle \vDash \phi[\alpha_1, \ldots, \alpha_n]$. The game given by $X_{(\phi,A)}$ is an *A-Neeman* game.

Definition 3.2 $\diamond_{\mathrm{G}}$*: For each* Σ^2_2 *sentence,* ϕ, $V \vDash \phi$ *if and only if* $V^{\mathrm{Coll}(\omega_1,\mathbb{R})} \vDash \phi$. □

The principle, $\diamond_{\mathrm{G}}$, is a *generic* form of $\diamond$. The next theorem gives a connection between versions of Σ^2_2 absoluteness and determinancy specifically the determinacy of Neeman games. In this theorem it is the principle, $\diamond_{\mathrm{G}}$, which plays the role of CH in the theorem on Σ^2_1 absoluteness.

Theorem 3.3 *Suppose that there exists a proper class of supercompact cardinals. Let* Γ^∞ *be the set of all* $A \subseteq \mathbb{R}$ *such that* A *is universally Baire. Then the following are equivalent.*

(1) *For each* $A \in \Gamma^\infty$, $\mathrm{ZFC} + \diamond_{\mathrm{G}} \vdash_\Omega$ " All A-Neeman games are determined".
(2) *For each* $A \in \Gamma^\infty$, *for each* Σ^2_2*-formula* $\phi(x)$, *either* $\mathrm{ZFC} + \diamond_{\mathrm{G}} \vdash_\Omega \phi[A]$ *or* $\mathrm{ZFC} + \diamond_{\mathrm{G}} \vdash_\Omega (\neg\phi)[A]$. □

We note the following trivial lemma which simply connects the results here with the earlier "evidence" that Σ^2_2 absoluteness is possible; cf. the discussion after Theorem 2.5.

Lemma 3.4 *Suppose that there exists a proper class of inaccessible limits of Woodin cardinals and suppose that for each* Σ^2_2*-sentence* ϕ, *either* $\mathrm{ZFC} + \diamond_{\mathrm{G}} \vdash_\Omega \phi$ *or* $\mathrm{ZFC} + \diamond_{\mathrm{G}} \vdash_\Omega (\neg\phi)$. *Then for each* Σ^2_2 *sentence* ϕ *the following are equivalent:*

(1) $\mathrm{ZFC} + \diamond_{\mathrm{G}} \vdash_\Omega \phi$*;*
(2) $\mathrm{ZFC} + \mathrm{CH} + \phi$ *is* Ω*-consistent.* □

The next theorem suggests that Σ^2_2 absoluteness conditioned simply on $\diamond$ might actually follow from some large cardinal hypothesis. Such a theorem would certainly be a striking generalization of Theorem 1.1 and its proof might well yield fundamental new insights into the combinatorics of subsets of ω_1.

Theorem 3.5 *Suppose that there exists a proper class of supercompact cardinals. Let* Γ^∞ *be the set of all* $A \subseteq \mathbb{R}$ *such that* A *is universally Baire and suppose that for each* $A \in \Gamma^\infty$, $\mathrm{ZFC} \vdash_\Omega$ " All A-Neeman games are determined". *Then for each* $A \in \Gamma^\infty$, *for each* Σ^2_2*-formula* $\phi(x)$, *either* $\mathrm{ZFC} + \diamond \vdash_\Omega \phi[A]$ *or* $\mathrm{ZFC} + \diamond \vdash_\Omega (\neg\phi)[A]$. □

Given Theorem 3.5, the natural conjecture is that Theorem 3.3 holds with $\diamond_{\mathrm{G}}$ replaced by $\diamond$. The missing ingredient in proving such a conjecture seems to be a lack of information on the nature of definable winning strategies for Neeman games and more fundamentally on the lack of any genuine determinacy proofs whatsoever for Neeman games.

In an exploration of the combinatorial aspects of Neeman games it is useful to consider a wider class of games. This class we now define.

For each formula, $\phi(x_1, \ldots, x_n)$, and for each stationary set $S \subseteq \omega_1$ let Y_ϕ be the set of all $a \in \{0,1\}^{\omega_1}$ such that there exists a stationary set $S \subseteq \omega_1$ such that for all $\alpha_1 < \cdots < \alpha_n$ in S, $\langle H(\omega_1), a, \in\rangle \vDash \phi[\alpha_1, \ldots, \alpha_n]$. The game given by Y_ϕ is a *stationary Neeman game.*

Can some large cardinal hypothesis imply that all stationary Neeman games are determined? Given the impossibility of $\Sigma^2_2(\mathcal{I}_{\rm NS})$-absoluteness, modulo failure of the Ω Conjecture one would naturally conjecture that the answer is "no". This is simply because there is no apparent candidate for an absoluteness theorem which would correspond to the (provable) determinacy of all stationary Neeman games.

We define two games of length ω_1. The first is a Neeman game and the second is a stationary Neeman game. Rather than have the moves be from $\{0,1\}$ it is more convenient to have the moves be from $H(\omega_1)$.

The canonical function game: Player I plays $< a_\alpha : \alpha < \omega_1\rangle$ and Player II plays $< b_\alpha : \alpha < \omega_1\rangle$ subject to the rules: $a_{\alpha+1} \subset \alpha \times \alpha$ and b_α is a countable ordinal.

Player I wins if there exists a set $A \subset \omega_1 \times \omega_1$ such that A is a wellordering of ω_1 and such that there exists a closed unbounded set $C \subset \omega_1$ such that for all $\alpha \in C$: $a_{\alpha+1} = A \cap (\alpha \times \alpha)$ and $b_\alpha < \text{rank}(a_{\alpha+1})$.

The stationary canonical function game: Player I plays $< a_\alpha : \alpha < \omega_1\rangle$ and Player II plays $< b_\alpha : \alpha < \omega_1\rangle$ subject to the rules: $a_{\alpha+1} \subset \alpha \times \alpha$ and b_α is a countable ordinal.

Player I wins if there exists a set $A \subset \omega_1 \times \omega_1$ such that A is a wellordering of ω_1 and such that there exists a statationary set $S \subset \omega_1$ such that for all $\alpha \in S$: $a_{\alpha+1} = A \cap (\alpha \times \alpha)$ and $b_\alpha < \text{rank}(a_{\alpha+1})$.

In models where L-like condensation principles hold these games are easily seen to be determined.

Lemma 3.6 *Suppose $\diamond$ holds. Then Player II has a winning strategy in the canonical function game.* □

Lemma 3.7 *Suppose $\diamond^+$ holds. Then Player II has a winning strategy in the stationary canonical function game.* □

In contrast to the previous lemma, the following theorem shows that it is consistent that Player I has a winning strategy in the stationary canonical function game, at least if fairly strong large cardinal hypotheses are assumed to be consistent.

Theorem 3.8 *Suppose there is a huge cardinal. Then there is a partial order, $\mathbb{P}$, such that in $V^{\mathbb{P}}$, Player I has a winning strategy in the stationary canonical function game.* □

These two results strongly suggest that no large cardinal hypothesis can imply that the stationary canonical function game is determined. In fact from consistency of a relatively weak large cardinal hypothesis, one does obtain the consistency that the stationary canonical function game is not determined. Note that if the stationary canonical function game is not determined then every function, $f : \omega_1 \to \omega_1$, is

bounded by a canonical function on a stationary set and so the consistency of *some* large cardinal hypothesis is necessary.

Theorem 3.9 *Suppose there is a measurable cardinal. Then there is a partial order, $\mathbb{P}$, such that in $V^{\mathbb{P}}$ the stationary canonical function game is not determined.*□

There are many open problems about the canonical function games. Here are several.

1. *Is it consistent that Player I has a winning strategy in the canonical function game?*
2. *Is it consistent that Player II does not have a winning strategy in the canonical function game?*
3. *Is it consistent that Player I has a winning strategy in the stationary canonical function game on each stationary set?*
4. *How strong is the assertion that Player I has a winning strategy in the stationary canonical function game?*

For each formula, $\phi(x_1, \ldots, x_n)$, for each sequence

$$\mathcal{S} = \langle S_\alpha : \alpha < \omega_1 \rangle$$

of pairwise disjoint stationary subsets of ω_1 and that $A \subseteq \mathbb{R}$, let $Y^{\mathcal{S}}_{(\phi,A)}$ be the set of all $a \in \{0,1\}^{\omega_1}$ such that there exists a stationary set $S \subseteq \omega_1$ such that for all $\alpha_1 < \cdots < \alpha_n$ in S,

$$\langle H(\omega_1), a, A, \in \rangle \vDash \phi[\alpha_1, \ldots, \alpha_n],$$

and such that $S \cap S_\alpha$ is stationary for all $\alpha < \omega_1$.

Theorem 3.10 *Suppose that there exists a proper class of supercompact cardinals. Let Γ^∞ be the set of all $A \subseteq \mathbb{R}$ such that A is universally Baire.*

Suppose that $A \in \Gamma^\infty$, $\phi(x_1, \ldots, x_n)$ is a formula and that

$$\mathrm{ZFC} \vdash_\Omega \text{“ The Neeman game } X_{(\phi,A)} \text{ is determined”}.$$

Then either:

(1) $\mathrm{ZFC} \vdash_\Omega$ “ I wins the game $X_{(\phi,A)}$”, *or;*
(2) $\mathrm{ZFC} \vdash_\Omega$ “ For all $\mathcal{S}$, II wins the game $Y^{\mathcal{S}}_{(\phi,A)}$”. □

The determinacy hypothesis: *All Neeman games are determined*; is relatively weak in consistency strength (the consistency strength is at most that of the existence of a Woodin cardinal which is a limit of Woodin cardinals). However the determinacy hypothesis:

For each formula ϕ, either Player I wins the game X_ϕ, or for each sequence,

$$\mathcal{S} = \langle S_\alpha : \alpha < \omega_1 \rangle,$$

of pairwise disjoint stationary subsets of ω_1, Player II wins $Y^{\mathcal{S}}_{(\phi,\emptyset)}$;

seems plausibly very strong.

References

[1] Qi Feng, Menachem Magidor, and Hugh Woodin. Universally Baire sets of reals. In *Set Theory of the Continuum (Berkeley, CA, 1989)*, 203–242. Springer, New York, 1992.

[2] Itay Neeman. *Long games.* Monograph, preprint. August, 2002.

[3] Itay Neeman. *Oberwolfach lecture.* January, 2002.

[4] J. Steel. *Determinacy plus* CH *implies* Π^2_1 *has the Scale Property.* preprint.

[5] W. Hugh Woodin. *The Axiom of Determinacy, forcing axioms, and the nonstationary ideal.* Walter de Gruyter & Co., Berlin, 1999.

ICM 2002 · Vol. I · 525–538

Evolution of Curves and Surfaces by Mean Curvature*

Brian White†

Abstract

This article describes the mean curvature flow, some of the discoveries that have been made about it, and some unresolved questions.

2000 Mathematics Subject Classification: 53C44.
Keywords and Phrases: Mean curvature flow, Singularities.

1. Introduction

Traditionally, differential geometry has been the study of curved spaces or shapes in which, for the most part, time did not play a role. In the last few decades, on the other hand, geometers have made great strides in understanding shapes that evolve in time. There are many processes by which a curve or surface can evolve, but among them one is arguably the most natural: the mean curvature flow. This article describes the flow, some of the discoveries that have been made about it, and some unresolved questions.

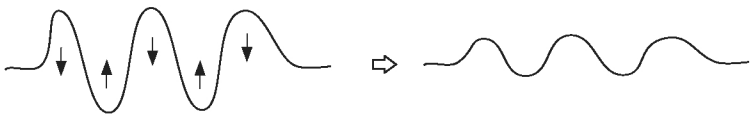
2. Curve-shortening flow

The simplest case is that of curves in the plane. Here the flow is usually called the "curvature flow" or the "curve-shortening flow". Consider a smooth simple closed curve in the plane, and let each point move with a velocity equal to the curvature vector at that point. What happens to the curve?

The evolution has several basic properties. First, it makes the curve smoother. Consider a portion of a bumpy curve as in figure 1(a). The portions that stick up move down and the portions that stick down move up, so the curve becomes smoother or less bumpy as in figure 1(b). The partial differential equation for the motion is a parabolic or heat-type equation, and such smoothing is a general feature

*The preparation of this article was partially funded by NSF grant DMS 0104049.
†Department of Mathematics, Stanford University, Stanford, CA 94305, USA.
E-mail: white@math.stanford.edu, Webpage: http://math.stanford.edu/~white

of solutions to such equations. Thus, for example, even if the initial curve is only C^2, as it starts moving it immediately becomes C^∞ and indeed real analytic.



1(a) 1(b)

Figure 1: Smoothing

However, there is an important caveat: the smoothing may only be for a short time. If the curve is C^2 at time 0, it will be real-analytic for times t in some interval $(0, \epsilon)$. But because the equation of motion is nonlinear, the general theory of parabolic equations does not preclude later singularities. And indeed, as we shall see, any curve must eventually become singular under the curvature flow.

The simplest closed curve is of course a circle. The flow clearly preserves the symmetry, so in this case it is easy to solve the equation of motion explicitly. One finds that a circle of radius of radius r at time 0 shrinks to a circle of radius $\sqrt{r^2 - 2t}$ at time t, so that time $t = r^2/2$ the circle has collapsed to a point and thereby become singular.

The second fundamental property is that arclength decreases. The proof is as follows. For any evolution of curves,

$$\frac{d}{dt}(\text{length}) = -\int \mathbf{k} \cdot \mathbf{v}\, ds,$$

where $\mathbf{k}$ is the curvature vector, $\mathbf{v}$ is the velocity, and ds is arclength. For the curvature flow, $\mathbf{v} = \mathbf{k}$, so the right hand side of the equation is clearly negative. Indeed, the proof shows that this flow is, in a sense, the gradient flow for the arclength functional. Thus, roughly speaking, the curve evolves so as to reduce its arclength as rapidly as possible. This explains the name "curve-shortening flow", though many other flows also reduce arc-length.

The third property is that the flow is collision-free: two initially disjoint curves must remain disjoint. The idea of the proof is as follows. Suppose that two initially disjoint curves, one inside the other, eventually collide. At the first time T of contact, they must touch tangentially. At the point of tangency, the curvature of the inner curve is greater than or equal to the curvature of the outer curve. Suppose for simplicity that strict inequality holds. Then (at the point of tangency) the inner curve is moving inward faster than the outer curve is. But then at a slighly earlier time $T - \epsilon$, the curves would have to cross each other. But that contradicts the choice of T (as first time of contact), proving that contact can never occur.

This collision avoidance is a special case of the maximum principle for parabolic differential equations. The maximum principle also implies in the same way that an initially embedded curve must remain embedded.

The fourth fundamental property is that every curve Γ has a finite lifespan. To undertand why, consider a circle C that contains Γ in its interior. Let Γ and C both evolve. The circle collapses in a finite time. Since the curves remain disjoint, Γ must disappear before the circle collapses.

There is another nice way to see that a curve must become singular in a finite time:

Theorem 1. *If $A(t)$ is the area enclosed by the curve at time t, then $A'(t) = -2\pi$ until the curve becomes singular. Thus a singularity must develop within time $A(0)/2\pi$.*

Proof. For any evolution,

$$A'(t) = \int_{\Gamma(t)} \mathbf{v} \cdot \mathbf{n}\, ds$$

where $\Gamma(t)$ is the curve at time t, $\mathbf{v}$ is the velocity, $\mathbf{n}$ is the outward unit normal, and ds is arclength. For the curvature flow, $\mathbf{v} = \mathbf{k}$, so

$$A'(t) = \int_{\Gamma} \mathbf{k} \cdot \mathbf{n}\, ds$$

which equals -2π by the Gauss-Bonnet theorem. □

The first deep theorem about curvature flow was proved by Mike Gage and Richard Hamilton in 1986 [GH]:

Theorem 2. *Under the curvature flow, a convex curve remains convex and shrinks to a point. Furthermore, it becomes asymptotically circular: if the evolving curve is dilated to keep the enclosed area constant, then the rescaled curve converges to a circle.*

This theorem is often summarized by stating that convex curves shrink to round points.

The proof is too involved to describe here, but I will point out that the result is not at all obvious. Consider for example a long thin ellipse, with the major axis horizontal. The curvature is greater at the ends than at the top and bottom, so intuitively it should become rounder. But the ends are much farther from the center than the top and bottom are, so it is not clear that they all reach the center at the same time. Thus it is not obvious that the curve collapses to a point rather than a segment.

Indeed, there are natural flows that have all the above-mentioned basic properties of curvature flow but for which the Gage-Hamilton theorem fails. Consider for example a curve moving in the direction of the curvature vector but with speed equal to the cube root of the curvature. Under this flow, any ellipse remains an ellipse of the same eccentricity and thus does not become circular. For this flow, Ben Andrews [A1] has proved that any convex curve shrinks to an elliptical point. (See also [AST, SaT].) For other flows (e.g. if "cube root" is replaced by rth root for any $r > 3$), a convex curve must shrink to a point but in a very degenerate

way: if the evolving curve is dilated to keep the enclosed area constant, then length of the rescaled curve tends to infinity [A3]. More generally, Andrews has studied the existence and nonexistence of asymptotic shapes for convex curves for rather general classes of flows [A2, A3].

Shortly after Gage and Hamilton proved their theorem, Matt Grayson proved what is still perhaps the most beautiful theorem in the subject:

Theorem 3. [G1] *Under the curvature flow, embedded curves become convex and thus (by the Gage-Hamilton theorem) eventually shrink to round points.*

Again, the proof is too complicated to describe here, but let me indicate why the result is very surprising. Consider the annular region between a two concentric circles of radii and $r = 1$ and $R = 2$. Form a curve in this annular region by spiraling inward n times, and then back out n times to make a closed embedded curve. Figure 2 shows such a curve with $n = 3/2$, but think of n being very large, say 10^{100}. Recall that the curve exists for a time at most $R^2/2 = 2$. By Grayson's theorem, the curve manages, amazingly, to unwind itself and become convex in this limited time. Incidentally, notice that initially, except for two very small portions, the curve is not even moving fast: its curvature is no more than that of the inner circle.

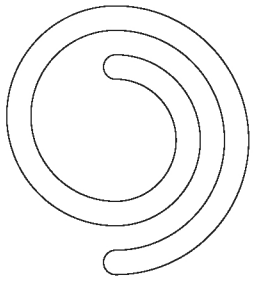


Figure 2: Spiral

As a corollary to Grayson's theorem, one gets an exact formula for the lifespan of any curve. Recall that the area enclosed by a curve decreases with constant speed -2π as long as the curve is smooth. By Grayson's theorem, the curve remains smooth until its area becomes 0. Thus the lifespan of any embedded curve must be exactly equal to the initial area divided by 2π.

Grayson later generalized his theorem by proving that a closed curve moving on a compact surface by curvature flow must either collapse to a round point in a finite time or else converge to a simple closed geodesic as $t \to \infty$ [G2].

3. Mean curvature flow for surfaces

We now leave curves and consider instead moving surfaces. Recall that at each point of an n-dimensional hypersurface in $\mathbf{R}^{n+1}$, there are n-principal curvatures

$\kappa_1, \kappa_2, \ldots, \kappa_n$ (given a choice $\mathbf{n}$ of unit normal). Their sum h is the scalar mean curvature, and the product $H = h\mathbf{n}$ of the scalar mean curvature and the unit normal is the mean curvature vector. The mean curvature vector does not depend on choice of normal since replacing $\mathbf{n}$ by $-\mathbf{n}$ also changes the sign of the scalar mean curvature. In the mean curvature flow, a hypersurface evolves so that its velocity at each point is equal to the mean curvature vector at that point.

The basic properties of curvature flow also hold for mean curvature flow:

(1) Surfaces become smoother for a short time.
(2) The area decreases. Indeed, mean curvature flow may be regarded as gradient flow for the area functional.
(3) Disjoint surfaces remain disjoint, and embedded surfaces remain embedded.
(4) Compact surfaces have limited lifespans.

The analog of the Gage-Hamilon theorem also holds, as Gerhard Huisken [H1] proved:

Theorem 4. *For $n \geq 2$, an n-dimensional compact convex surface in $\mathbf{R}^{n+1}$ must shrink to a round point.*

Oddly enough, Huisken's proof does not apply to the case of curves ($n = 1$) considered by Gage and Hamilton. Huisken's proof shows that the asymptotic shape is totally umbilic: at each point x, the principal curvatures are all equal (though *a priori* they may vary from point to point). For $n \geq 2$, the only totally umbilic surfaces are spheres, but for $n = 1$, the condition is vacuous.

The analog of Grayson's theorem, however, is false for surfaces. Consider for example two spheres joined by a long thin tube. The spheres and the tube both shrink, but the mean curvature along the tube is much higher than on the spheres, so the middle of the tube collapses down to a point, forming a singularity. The surface then separates into two components, which eventually become convex and collapse to round points.

Thus, unlike a curve, a surface can develop singularities before it shrinks away. This raises various questions:

(1) How do singularities affect the subsequent evolution of the surface?
(2) How large can the set of singularities be?
(3) What is the nature of the singularities? What does the surface look like near a singular point?

In the rest of this article I will describe some partial answers to these questions.

A great many results about mean curvature flow have been proved using only techniques of classical differential geometry and partial differential equations. However, most of the proofs are valid only until the time that singularities first occur. Once singularities form, the equation for the flow does not even make sense classically, so analyzing the flow after a singularity seems to require other techniques.

Fortunately, using non-classical techniques, namely the geometric measure theory of varifolds and/or the theory of viscosity or level-set solutions, one can define notions of weak solutions for mean curvature flow and one can prove existence of a

solution (with a given initial surface) up until the time that the surface disappears.

The definitions are somewhat involved and will not be given here. The different definitions are equivalent to each other (and to the classical definition) until singularities form, but are not completely equivalent in general. For the purposes of this article, the reader may simply accept that there is a good way to define mean curvature flow of possibly singular surfaces and to prove existence theorems. The notion of mean curvature flow most appropriate to this article is Ilmanen's "enhanced Brakke flow of varifolds" [I].

4. Non-uniqueness or fattening

If a surface is initially smooth, classical partial differential equations imply that there is a unique solution of the evolution equation until singularities form. However, once a singularity forms, the classical uniqueness theorems do not apply. In the early 1990's various researchers, including De Giorgi, Evans and Spruck, and Chen, Giga, and Goto, asked whether uniqueness could in fact break down after singularity formation. They already knew that uniqueness did fail for certain initially singular surfaces, but they did not know whether an initially regular surface could later develop singularities that would result in non-uniqueness.

A technical aside: the above-named people did not phrase the question in terms of uniqueness but rather in terms of "fattening". They were all using a level set or viscosity formulation of mean curvature flow, in which solutions are almost by definition unique. But non-uniqueness of the enhanced varifold solutions corresponds to fattening of the viscosity solution in the following sense. If a single initial surface M gives rise to different enhanced varifold solutions $M_t^1, M_t^2, \ldots, M_t^k$, then the viscosity solution "surface" at time t will consist of the various M_t^i's together with all the points in between. Thus if $k > 1$, the viscosity surface will in fact have an interior. Since the surface was initially infinitely thin, in developing an interior it has thereby "fattened".

Recently Tom Ilmanen and I settled this question [IW]:

Theorem 5. *There is a compact smooth embedded surface in* $\mathbf{R}^3$ *for which uniqueness of (enhanced varifold solutions of) mean curvature evolution fails. Equivalently, the viscosity (or level set) solution fattens.*

The idea of the proof is as follows. Consider a solid torus of revolution about the z-axis centered at the origin, a ball centered at the origin that is disjoint from the torus, and n radial segments in the xy-plane joining the ball to the torus. Call their union W. Now consider a nested one-parameter family of smooth surfaces M^ϵ $(0 < \epsilon < 1)$ as follows. When ϵ is small, the surface should be a smoothed version of the set of points at distance ϵ from W. This M^ϵ looks like a wheel with n spokes. The portion of the xy-plane that is not contained in M^ϵ has n simply-connected components, which we regard as holes between the spokes of the wheel. As ϵ increases, the spokes get thicker and the holes between the spokes get smaller. When ϵ is close to 1, the holes should be very small, and near each hole the surface

should resemble a thin vertical tube.

Now let M^ϵ flow by mean curvature. If ϵ is small, the spokes are very thin and will quickly pinch off, separating the surface into a sphere and a torus. If ϵ is large, the holes between the spokes are very small and will quickly pinch off, so the surface becomes (topologically) a sphere. By a continuity argument, there is an intermediate ϵ such that both pinches occur simultaneously.

For this particular ϵ, we claim that the simultaneous pinching immediately results in non-uniqueness, at least if n is sufficiently large. Indeed, at the moment of simultaneous pinching, the surface will resemble a sideways figure 8 curve revolved around the z-axis as indicated in figure 3(a). Of course the surface will not be fully rotationally symmetric about the axis, but it will have n-fold rotational symmetry, and here I will be imprecise and proceed as though it were rotationally symmetric.

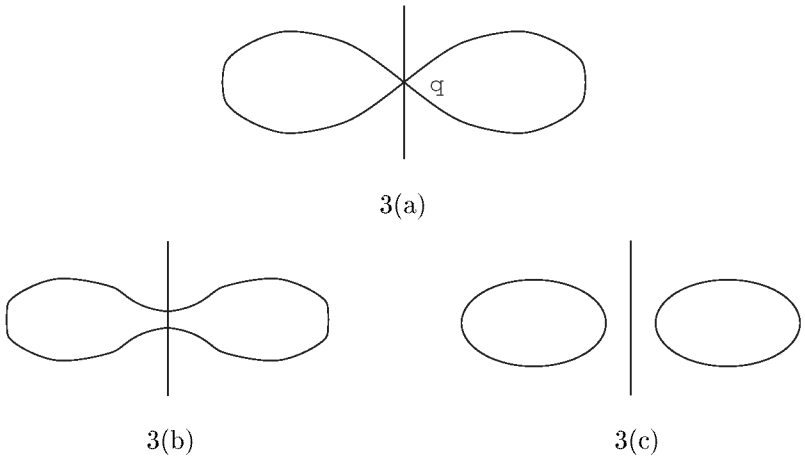


Figure 3: Non-uniqueness

There is necessarily one evolution in which the surface then becomes topologically a sphere as in figure 3(b). If the angle θ in figure 6 is sufficiently small, there is also another evolution, in which the surface detaches itself from the z-axis and thereby becomes a torus as in figure 3(c).

One can show that as $n \to \infty$, the angle θ tends to 0. Thus if n is large enough, the angle will be very small and both evolutions will occur.

The proof unfortunately does not give any bound on how large an n is required. Numerical evidence [AIT] suggests that $n = 4$ suffices. The case $n = 2$ seems to be borderline and the case $n = 3$ has not been investigated numerically.

However, the argument completely breaks down for $n < 2$. Indeed, I would conjecture that if the initial surface is a smooth embedded sphere or torus, then uniqueness must hold.

It is desirable to know natural conditions on the initial surface that guarantee

uniqueness. As just mentioned, genus ≤ 1 may be such a condition. Mean convexity (described below) and star-shapedness are known to guarantee uniqueness. The latter is interesting because the surface will typically cease to be star-shaped after a finite time, but its initially starry shape continues to ensure uniqueness [So].

Fortunately, uniqueness is known to hold generically in a rather strong sense. If a family of hypersurfaces foliate an open set in $\mathbf{R}^{n+1}$, then uniqueness will hold for all except countably many of the leaves. Of course any smooth embedded surface is a leaf of such a foliation, so by perturbing the surface slightly, we get a surface for which uniqueness holds.

5. The size of singular sets

For general initial surfaces, our knowledge about singular sets is very limited. Concerning the size of the singular set, Tom Ilmanen [I], building on earlier deep work of Ken Brakke [BK], proved the following theorem. (See also [ES IV].)

Theorem 6. *For almost every initial hypersurface M_0 and for almost every time t, the surface M_t is smooth almost everywhere.*

This theorem reminds me of Kurt Friedrichs, who used to say that he did not like measure theory because when you do measure theory, you have to say "almost everywhere" almost everywhere.

Aside from objections Friedrichs might have had, the theorem is unsatisfactory in that a much stronger statement should be true. But it is a tremendous acheivement and it is the best result to date for general initial surfaces.

However, for some classes of initial surfaces, we now have a much better understanding of singularities. In particular, this is the case when the initial surface is **mean-convex**. The rest of this article is about such surfaces. For simplicity of language, only two-dimensional surfaces in $\mathbf{R}^3$ will be discussed, but the results all have analogs for n-dimensional surfaces in $\mathbf{R}^{n+1}$ or, more generally, in $(n+1)$-dimensional riemannian manifolds.

Consider a compact surface M embedded in $\mathbf{R}^3$ and bounding a region Ω. The surface is said to be "mean-convex" if the mean curvature vector at each point is a nonnegative multiple of the inward unit normal (that is, the normal that points into Ω.) This is equivalent to saying that under the mean curvature flow, M immediately moves into Ω. Mean convexity is a very natural condition for mean curvature flow:

(1) If a surface is initially mean convex, then it remains mean convex as it evolves.
(2) Uniqueness (or non-fattening) holds for mean convex surfaces.

Mean convexity, although a strong condition, does not preclude interesting singularity formation. For example, one can connect two spheres by a thin tube as described earlier in such a way that the resulting surface is mean convex. Thus neck pinch singularities do occur for some mean convex surfaces.

Theorem 7. [W1] *A mean convex surface evolving by mean curvature flow in $\mathbf{R}^3$ must be completely smooth (i.e., with no singularities) at almost all times, and*

at no time can the singular set be more than 1 *dimensional.*

This theorem is in some ways optimal. For example, consider a torus of revolution bounding a region Ω. If the torus is thin enough, it will be mean convex. Because the symmetry is preserved and because the surface always remains in Ω, it can only collapse to a circle. Thus at the time of collapse, the singular set is one-dimensional.

However, in other ways the result is probably not optimal. In particular, the result should hold without the mean convexity hypothesis, and singularities should occur at only finitely many times. Indeed, I would conjecture that at each time, the surface can only have finitely many singularities unless one or more connected components have collapsed to curves. That is, the surface should consist of finitely many connected components, each of which either is a curve or has only finitely many singularities.

6. Nature of mean-convex singularities

Recall that when a mean convex surface evolves, it starts moving inward. Because mean convexity is preserved, it must continue to move inward. Consequently, the surface at any time lies strictly inside the region bounded by the surface at any previous time. Since the motion is continuous and since the surface collapses in a finite time, this implies that region Ω bounded by M_0 is the disjoint union of the M_t's for $t > 0$. It is convenient and suggestive to speak of the M_t's forming a foliation of Ω, although it is not quite a foliation in the usual sense because some of the leaves are singular. Figure 4 shows the foliation when the initial surface is two spheres joined by a thin tube. (The entire foliation is rotationally symmetric about an axis, so suffices to show the intersection of the foliation with a plane containing that axis.)

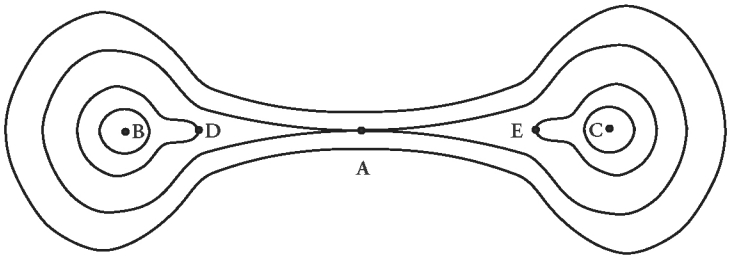


Figure 4: Foliation by evolving mean-convex surfaces

Theorem 8. [W2] *Consider a mean convex surface M_t $(t \geq 0)$ in $\mathbf{R}^3$ evolving by mean curvature flow. Let p be any singular point in the region Ω bounded by the initial surface. If we dilate about p by a factor λ and then then let $\lambda \to \infty$, the*

dilated foliation must converge subsequentially to a foliation of $\mathbf{R}^3$ *consisting of*

(1) *parallel planes, or*
(2) *concentric spheres, or*
(3) *coaxial cylinders.*

Let us call such a subsequential limit a **tangent foliation** at p. The tangent foliation consists of parallel planes if and only if p is a regular point (i.e., p has a neighborhood U such that the $M_t \cap U$ smoothly foliate U.) A tangent foliation of concentric spheres corresponds to M_t (or a component of M_t) becoming convex and collapsing as in Huisken's theorem (theorem 4) to the round point p.

In figure 4, the tangent foliation at the "neck pinch" point A is a foliation by coaxial cylinders. The tangent foliations at points B and C are by concentric spheres. All other points are regular points and thus give rise to tangent foliations consisting of parallel planes.

Incidentally, in cases (1) and (2), the tangent foliation is unique. That is, we have convergence and not just subsequential convergence in the statement of theorem 8. However, this is not known in case (3). If one tangent foliation at p consists of cylinders, then so does any other tangent folation at p [S]. But it is conceivable that different sequences of λ's tending to infinity could give rise to cylindrical foliations with different axes of rotational symmetries. Whether this can actually happen is a major unsolved problem, exactly analogous to the long-standing uniqueness of tangent cone problem is minimal surface theory.

Although the tangent foliation at a singular point carries much information about the singularity, there are features that it misses. For example, consider the neck pinch in figure 4, located at the point A. At a time just after the neck pinch, the two points D and E on the surface that are nearest to A have very large mean curvature and are therefore moving away from A very rapidly. However, such behavior cannot be seen in tangent foliations: the tangent foliation at any point near A consists of parallel planes, and the tangent foliation at A consists of coaxial cylinders.

To capture behavior such as the rapid motion away from A, rather than dilating about a fixed point as in theorem 8, one needs to track a moving point.

Theorem 9. *Consider a mean convex surface* M_t $(t \geq 0)$ *in* $\mathbf{R}^3$ *evolving by mean curvature flow. Let* p_i *be a sequence of points converging to a point* p *in the region bounded by* M_0*, and let* λ_i *be a sequence of numbers tending to infinity. Translate the* M_t*'s by* $-p_i$ *and then dilate by* λ_i*. Then the resulting sequence of foliations must converge subsequentially to a foliation of* $\mathbf{R}^3$ *by one of the following:*

(1) *compact convex sets, or*
(2) *coaxial cylinders, or*
(3) *parallel planes, or*
(4) *non-compact strictly convex surfaces, none of which are singular.*

The convergence is locally smooth away from the limit foliation's singular set (a point in case (1)*, a line in case* (2)*, and the empty set in the other two cases.)*

A foliation obtained in this way is called a **blow-up foliation** at p. Of course if all the p_i's are equal to p, we get a tangent foliation at p.

It follows from the smooth convergence that a blow-up foliation is invariant under mean-curvature flow. That is, if we let a leaf flow for a time t, the result will still be a leaf. Consequently, except for the case (3) of parallel planes (which do not move under the flow), given any two leaves, one will flow to the other in finite time. Thus we can index the leaves as M'_t in such a way that when M'_t flows for a time s, it becomes M'_{t+s}. In the cases (1) and (2) of compact leaves and cylindrical leaves, the indexing interval may be taken to be $(-\infty, 0]$. In case (3), the indexing interval is $(-\infty, \infty)$. Thus, except in the case of parallel planes, the blow-up foliation corresponds to a semi-eternal or eternal flow of convex sets that sweep out all of $\mathbf{R}^3$.

As pointed out earlier, blow-up foliations (1), (2), and (3) already occur as tangent foliations. (If (2) occurs as a tangent foliation, then the compact convex sets must all be spheres, but in general blow-up foliations, other compact convex sets might conceivably occur.)

Thus the new case is (4). To see how that case arises, consider in figure 4 a sequence of points E_i on the axis of rotational symmetry that converge to the neck pinch point A. Let h_i be the mean curvature at E_i of the leaf of the foliation that passes through E_i.

Now if we translate the foliation by $-E_i$ and then dilate by $\lambda_i = h_i$, then the dilated foliations converge to a blow-up foliation of $\mathbf{R}^3$ by convex non-compact surfaces. Each leaf qualitatively resembles a rotationally symmetric paraboloid $y = x^2 + z^2$. Furthermore, the leaves are all translates of each other. In other words, if we let one of the leaves evolve by mean curvature flow, then it simply translates with constant speed.

Incidentally, the same points E_i with different choices of λ_i's can give rise to different blow-up foliations. For if the dilation factors $\lambda_i \to \infty$ quickly compared to h_i (that is, if $\lambda_i/h_i \to \infty$), then the resulting blow-up foliation consists of parallel planes. If $\lambda_i \to \infty$ slowly compared to h_i (so that $\lambda_i/h_i \to 0$), then the resulting foliation consists of coaxial cylinders.

So what is the "right" choice of λ_i? In a way, it depends on what one wants to see. But this example does illustrate a general principle:

Theorem 10. [W2] *Let $p_i \in M_{t(i)}$ be a sequence of regular points converging to a singular point p. Translate $M_{t(i)}$ by $-p_i$ and the dilate by h_i (the mean curvature of $M_{t(i)}$ at p_i) to get a new surface M'_i. Then a subsequence of the M'_i will converge smoothly on bounded subset of $\mathbf{R}^3$ to a smooth strictly convex surface M'.*

Of course M' is one leaf of the corresponding blow-up foliation.

Theorems 8, 9, and 10 give a rather precise picture of the singular behavior, but they raise some problems that have not yet been answered:

(1) Classify all the eternal and semi-eternal mean-curvature evolutions of convex sets that sweep out all of $\mathbf{R}^3$.

(2) Classify those associated with blow-up foliations.

The strongest conjecture for (2) is that a blow-up foliation can only consist of planes, spheres, cylinders, or the (unique) rotationally symmetric translating surfaces.

Many more eternal and semi-eternal evolutions of convex sets are known to exist. For instance, given any three positive numbers a, b, and c, there is a semi-eternal evolution of compact convex sets, each of which is symmetric about the coordinate planes and one of which cuts off segments of lengths a, b, and c from the x, y, and z axes, respectively. (This can be proved by a slight modification of the proof given for example 3 in the "conclusions" section of [W2].) The case $a = b = c$ is that of concentric spheres, which of course do occur as a blow-up foliation. Whether the other cases occur as blow-up foliations is not known.

A very interesting open question in this connection is: must every eternal evolution of convex sets consist of leaves that move by translating? Tom Ilmanen has recently shown that there is a one parameter family of surfaces that evolve by translation. At one extreme is the rotationally symmetric one, which does occur in blow-up foliations. At the other extreme is the Cartesian product of a certain curve with $\mathbf{R}$:

$$\{(x, y, z) : y = -\ln \cos x, \quad -1 < x < 1\}.$$

This case does not occur in blow-up foliations ([W2].)

7. Further reading

Three distinct approaches have been very fruitful in investigating mean curvature: geometric measure theory, classical PDE, and the theory of level-set or viscosity solutions. These were pioneered in [BK]; [H1] and [GH]; and [ES] and [CGG] (see also [OS]), respectively. Surveys emphasizing the classical PDE approach may be found in [E1] and [H3]. A very readable and rather thorough introduction to the classical approach, including some new results (as well as some discussion of geometric measure theory), may be found in [E2]. An introduction to the geometric measure theory and viscosity approaches is included in [I]. See [G] for a more extensive introduction to the level set approach.

Theorems 7, 8, 9, and 10 about mean convex surfaces are from my papers [W1] and [W2]. These papers rely strongly on earlier work, for instance on Brakke's regularity theorem and on Huisken's monotonicity formula. Huisken proved theorem 8 much earlier under a hypothesis about the rate at which the curvature blows up [H2]. Huisken and Sinestrari [HS 1,2] independently proved results very similar to theorems 8, 9, and 10, but only up to the first occurrence of singularities.

Much of the current interest in curvature flows stems from Hamilton's spectacular work on the Ricci flow. For survey articles about Ricci flow, see [CC] and [Ha]. For discussions of some other interesting geometric flows, see the articles by Andews [A4] and by Bray [BH] in these Proceedings.

References

[A1] B. Andrews, Contraction of convex hypersurfaces by their affine normal, *J. Differential Geometry,* 43 (1996), 207–230.

[A2] ———, Evolving convex curves, *Calc. Var. Partial Differential Equations,* 7 (1998), 315–371.

[A3] ———, Non-convergence and instability in the asymptotic behavior of curves evolving by curvature, *Comm. Anal. Geom.,* 10 (2002), 409–449.

[A4] ———, Positively curved surfaces in the three-sphere, Proceedings of the International Congress of Mathematicians, vol. II (Beijing, 2002), Higher Education Press, Beijing, 2002, 221–230.

[AIC] S. Angenent, T. Ilmanen, and D. L. Chopp, A computed example of nonuniqueness of mean curvature flow in $\mathbf{R}^3$, *Comm. Partial Differential Equations,* 20 (1995), no. 11-12, 1937–1958.

[AST] S. Angenent, G. Sapiro, and A. Tannenbaum, On the affine heat equation for non-convex curves, *J. Amer. Math. Soc.,* 11 (1998), no. 3, 601–634.

[BK] K. Brakke, *The motion of a surface by its mean curvature*, Princeton U. Press, 1978.

[BH] H. Bray, Black holes and the penrose inequality in general relativity, Proceedings of the International Congress of Mathematicians, vol. II (Beijing, 2002), Higher Education Press, Beijingm, 2002, 257–371.

[CC] H.-D. Cao and B. Chow, Recent developments on the Ricci flow, *Bull. Amer. Math. Soc. (N.S.),* 36 (1999), no. 1, 59–74.

[CGG] Y. G. Chen, Y. Giga, and S. Goto, Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations, *J. Diff. Geom.,* 33 (1991), 749–786.

[E1] K. Ecker, Lectures on geometric evolution equations, Instructional Workshop on analysis and geometry, part II (Canberra, 1995), Austral. Nat. Univ., Canberra, 1996, 79–107.

[E2] ———, Lectures on regularity for mean curvature flow, preprint (based on lectures given 2000-2001 at Universität Freiburg).

[ES] L. C. Evans and J. Spruck, Motion of level sets by mean curvature I, *J. Diff. Geom.,* 33 (1991), no. 3, 635–681; II, *Trans. Amer. Math. Soc.,* 330 (1992), no. 1, 321–332; III, *J. Geom. Anal.,* 2 (1992), no. 2, 121–150; IV, *J. Geom. Anal.,* 5 (1995), no. 1, 77–114.

[GH] M. Gage and R. Hamilton, The heat equation shrinking convex plane curves, *J. Differential Geom.,* 23 (1986), 417–491.

[G] Y. Giga, *Surface evolution equations – a level set method*, Hokkaido U. Tech. Report Series in Math.,, vol. 71, Dept. of Math., Hokkaido Univ., Sapporo 060-0810 Japan, 2002.

[Gr1] M. Grayson, The heat equation shrinks embedded plane curves to round points, *J. Differential Geom.,* 26 (1987), 285–314.

[Gr2] ———, Shortening embedded curves, *Ann. of Math. (2),* 129 (1989), no. 1, 71–111.

[Ha] R. Hamilton, The formation of singularities in the Ricci flow, Surveys in differential geometry, Vol. II (Cambridge, MA, 1993), Internat. Press, Cambridge, MA, 1995, 7–136.

[H1] G. Huisken, Flow by mean curvature of convex surfaces into spheres, *J. Diff. Geom.,* 20 (1984), 237–266.

[H2] ———, Local and global behavior of hypersurfaces moving by mean curvature, *Proc. Symp. Pure Math.,* 54 (1993), Amer. Math. Soc., 175–191.

[H3] ———, Lectures on geometric evolution equations, Tsing Hua lectures on geometry and analysis (Hsinhcu, 1990–1991), Internat. Press, Cambridge, MA, 1997, 117–143.

[HS1] G. Huisken and C. Sinestrari, Mean curvature flow singularities for mean convex surfaces, *Calc. Var. Partial Differential Equations,* 8 (1999), 1–14.

[HS2] ———, Convexity estimates for mean curvature flow and singularities of mean convex surfaces, *Acta Math.,* 183 (1999), 45–70.

[I] T. Ilmanen, Elliptic regularization and partial regularity for motion by mean curvature, *Memoirs of the AMS,* 108 (1994).

[IW] T. Ilmanen and B. White, Non-uniqueness for mean curvature flow of an initially smooth compact surface, in preparation.

[OS] S. Osher and J. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations, *J. Comput. Phys.,* 79 (1988), 12–49.

[SaT] G. Sapiro and A. Tannenbaum, On affine plane curvature evolution, *J. Funct. Anal.,* 119, no. 1, 79–120.

[So] H. M. Soner, Motion of a set by the curvature of its boundary, *J. Differential Equations,* 101 (1993), no. 2, 313–372.

[St1] A. Stone, A density function and the structure of singularities of the mean curvature flow, *Calc. Var. Partial Differential Equations,* 2 (1994), 443–480.

[W1] B. White, The size of the singular set in mean curvature flow of mean convex surfaces, *J. Amer. Math. Soc.,* 13 (2000), no. 3, 665–695.

[W2] ———, The nature of singularities in mean curvature flow of mean convex surfaces, *J. Amer. Math. Soc.* 16 (2003), 123–138.

ICM 2002 · Vol. I · 539–546

On Hrushovski's Proof of the Manin-Mumford Conjecture

Richard Pink* Damian Roessler†

Abstract

The Manin-Mumford conjecture in characteristic zero was first proved by Raynaud. Later, Hrushovski gave a different proof using model theory. His main result from model theory, when applied to abelian varieties, can be rephrased in terms of algebraic geometry. In this paper we prove that intervening result using classical algebraic geometry alone. Altogether, this yields a new proof of the Manin-Mumford conjecture using only classical algebraic geometry.

2000 Mathematics Subject Classification: 14K12.
Keywords and Phrases: Torsion points, Abelian varieties, Manin-Mumford conjecture.

1. Introduction

The Manin-Mumford conjecture is the following statement.

Theorem 1.1 *Let A be an abelian variety defined over $\overline{\mathbb{Q}}$ and X a closed subvariety of A. Denote by* $\mathrm{Tor}(A)$ *the set of torsion points of A. Then*

$$X \cap \mathrm{Tor}(A) = \bigcup_{i \in I} X_i \cap \mathrm{Tor}(A),$$

where I is a finite set and each X_i is the translate by an element of A of an abelian subvariety of A, immersed in X.

This conjecture has a long history of proofs. A first partial proof was given by Bogomolov in [2], who proved the statement when $\mathrm{Tor}(A)$ is replaced by its

*Department of Mathematics, ETH-Zentrum, CH-8092 Zürich, Switzerland. E-mail: pink@math.ethz.ch
†Department of Mathematics, ETH-Zentrum, CH-8092 Zürich, Switzerland. E-mail: roessler@math.ethz.ch

ℓ-primary part for a prime number ℓ; he applies results of Serre, Tate and Raynaud (see [11]) on the existence of Hodge-Tate module structures on the Tate module of abelian varieties over discrete valuation rings. A full proof of the conjecture was then given by Raynaud in [9] (see [8] for the case $\dim(X) = 1$); his proof follows from a study of the reduction of A modulo p^2. A third and full proof of the conjecture was given by Hrushovski in [4]; he uses Weil's theorem on the characteristic polynomial of the Frobenius morphism on abelian varieties over finite fields in conjunction with the model theory (of mathematical logic) of the theory of fields with a distinguished automorphism. A fourth proof of the conjecture was given by Ullmo and Zhang (based on ideas of Szpiro) in [13] and [15] and goes via diophantine approximation and Arakelov theory. They actually prove a more general conjecture of Bogomolov which generalizes the Manin-Mumford conjecture.

This article was inspired by Hrushovski's proof. The bulk of Hrushovski's proof lies in the model theory part. It culminates in a result which, when applied to the special case of abelian varieties and stripped of model theoretic terminology, is essentially Theorem 2.1 below. In Section 2 we prove Theorem 2.1 with classical algebraic geometry alone. Neither scheme theory, nor Arakelov theory, nor mathematical logic are used. In section 3, for the sake of completeness, we show how to apply Theorem 2.1 to prove the Manin-Mumford conjecture.

In a subsequent article (cf. [7]), we shall consider the analogue of 2.1 for varieties over function fields of characteristic $p > 0$.

2. Hrushovski's theorem for abelian varieties

Let K be an algebraically closed field of characteristic zero, endowed with an automorphism σ. Let A be an abelian variety over K and X a closed subvariety of A. For ease of notation, we use the language of classical algebraic geometry; thus A and X denote the respective sets of K-valued points. We assume that $X \subset A$ are defined already over the fixed field K^σ. The automorphism of A induced by σ is again denoted by σ. Let $P(T) \in \mathbb{Z}[T]$ be a monic polynomial with integral coefficients. In [4, Cor. 4.1.13, p.90], Hrushovski proves the generalisation of the following theorem to semi-abelian varieties.

Theorem 2.1 *Let Γ denote the kernel of the homomorphism $P(\sigma) : A \to A$. Assume that no complex root of P is root of unity. Then*

$$X \cap \Gamma = \bigcup_{i \in I} X_i \cap \Gamma,$$

where I is a finite set and each X_i is the translate by an element of A of an abelian subvariety of A, immersed in X.

Remark If roots of unity are not excluded, the group Γ becomes too large for such a result. For example, if $T^m - 1$ divides P, all points of A over the fixed field K^{σ^m} of σ^m are contained in Γ.

Proof Write $P(T) = \sum_{i=0}^{n} a_i T^i$ with $a_i \in \mathbb{Z}$ and $a_n = 1$. Let F be the endomorphism of A^n defined by the matrix

$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & \dots & -a_{n-1} \end{pmatrix}$$

which is the companion matrix of the polynomial P, and note that $P(F) = 0$. In the obvious way σ induces an automorphism of A^n, denoted again by σ, that sends X^n to itself. Let Δ denote the kernel of the homomorphism $F - \sigma : A^n \to A^n$. By construction, there is a canonical bijection

$$\Gamma \to \Delta, \ x \mapsto (x, \sigma(x), \dots, \sigma^{n-1}(x)).$$

Since $\sigma(X) = X$, this induces a bijection

$$X \cap \Gamma \to X^n \cap \Delta.$$

Its inverse is given by the projection to the first factor $A^n \to A$. Clearly, we are now reduced to the following theorem, applied to $X^n \subset A^n$ in place of $X \subset A$.

Theorem 2.2 *Let $F : A \to A$ be an algebraic endomorphism that commutes with σ and that satisfies $P(F) = 0$. Let Δ denote the kernel of the homomorphism $F - \sigma : A \to A$. Assume that no complex root of P is root of unity. Then*

$$X \cap \Delta = \bigcup_{i \in I} X_i \cap \Delta,$$

where I is a finite set and each X_i is the translate by an element of A of an abelian subvariety of A, immersed in X.

Remark 2.3 If K is algebraic over the fixed field K^σ, every element $a \in \Delta$ satisfies $F^m(a) = \sigma^m(a) = a$ for some $m \geq 1$. In other words, we have $a \in \mathrm{Ker}(F^m - \mathrm{id})$. The assumptions on F and P imply that $F^m - \mathrm{id}$ is an isogeny; hence a is a torsion element. It follows that Δ is a torsion subgroup, and the theorem follows from the Manin-Mumford conjecture in this case. However, the scope of the above theorem is somewhat wider, and the Manin-Mumford conjecture will be deduced from it.

Proof Let Y be the Zariski closure of $X \cap \Delta$. We claim that $\sigma(Y) = F(Y) = Y$. To see this, note first that σ commutes with F, and so $\sigma(\Delta) = \Delta$. By assumption we have $\sigma(X) = X$; hence $\sigma(X \cap \Delta) = X \cap \Delta$. Since $\sigma : A \to A$ comes from an automorphism of the underlying field, it is a homeomorphism for the Zariski topology, so we have $\sigma(Y) = Y$. On the other hand, the maps σ and F coincide on Δ, which implies $F(X \cap \Delta) = X \cap \Delta$. As F is a proper algebraic morphism, we deduce that $F(Y) = Y$. Clearly, Theorem 2.2 is now reduced to the following theorem (see [3, Th. 3] for the case where F is the multiplication by an integer $n > 1$).

Theorem 2.4 *Let A be an abelian variety over an algebraically closed field of characteristic zero. Let $F : A \to A$ be an algebraic endomorphism none of whose eigenvalues on* $\mathrm{Lie}\, A$ *is a root of unity. Let Y be a closed subvariety of A satisfying $F(Y) = Y$. Then Y is a finite union of translates of abelian subvarieties of A.*

Proof We proceed by induction on the dimension d of A. For $d = 0$ the statement is obvious; hence we assume $d > 0$. Next observe that $Y \subset F(A)$. Thus if F is not surjective, we can replace A by $F(A)$ and Y by $F(A) \cap Y$, and are finished by induction. Thus we may assume that F is an isogeny.

Since F is proper and $F(Y) = Y$, every irreducible component of Y is the image under F of some irreducible component of Y. Since the set of these irreducible components is finite, it is therefore permuted by F. We fix such an irreducible component Z and an integer $m \geq 1$ such that $F^m(Z) = Z$.

Any power $F^{rm} : A \to A$ is an isogeny, and since $\mathrm{char}(K) = 0$, it is a separable isogeny. As a morphism of schemes it is therefore a finite étale Galois covering with Galois group $\mathrm{Ker}(F^{rm})$, acting by translations on A. The same follows for the induced covering $(F^{rm})^{-1}(Z) \to Z$. As Z is irreducible, the irreducible components of $(F^{rm})^{-1}(Z)$ are transitively permuted by $\mathrm{Ker}(F^{rm})$ and each of them has dimension $\dim(Z)$. Since $F^{rm}(Z) = Z$, we have $Z \subset (F^{rm})^{-1}(Z)$, and so Z itself is one of these irreducible components. Let G_r denote the stabilizer of Z in $\mathrm{Ker}(F^{rm})$. Then $F^{rm} : Z \to Z$ is generically a finite étale Galois covering with Galois group G_r.

We now distinguish two cases. Let $\mathrm{Stab}_A(Z)$ denote the translation stabilizer of Z in A, which is a closed algebraic subgroup of A.

Lemma 2.5 *If $|G_1| > 1$, then* $\dim(\mathrm{Stab}_A(Z)) > 0$.

Proof For any $r \geq 1$, the morphism $F^{rm} : Z \to Z$ is finite separable of degree $|G_r|$. Since degrees are multiplicative in composites, this degree is also equal to $|G_1|^r$. Thus if $|G_1| > 1$, we find that $|G_r|$ becomes arbitrarily large with r. Therefore $\mathrm{Stab}_A(Z)$ contains arbitrarily large finite subgroups, so it cannot be finite. **Q.E.D.**

Lemma 2.6 *If $|G_1| = 1$, then* $\dim(\mathrm{Stab}_A(Z)) > 0$ *or Z is a single point.*

Proof Since Z is irreducible, the assertion is obvious when $\dim(Z) = 0$. So we assume that $\dim(Z) > 0$. The assumption $|G_1| = 1$ implies that F^m induces a finite (separable) morphism $\varphi : Z \to Z$ of degree 1. The set of fixed points of any positive power φ^r is then $Z \cap \mathrm{Ker}(F^{rm} - \mathrm{id})$. On the other hand the assumptions on F and P imply that $F^{rm} - \mathrm{id}$ is an isogeny on A. Thus this fixed point set is finite for every $r \geq 1$; hence φ has infinite order.

Assume now that $\mathrm{Stab}_A(Z)$ is finite. By Ueno's theorem [12, Thm. 3.10] Z is then of general type, in the sense that any smooth projective variety birationally equivalent to Z is of general type (see [12, Def. 1.7]). But the group of birational automorphisms of any irreducible projective variety Z of general type is finite (see [6] and also [5, Th. 10.11]) which yields a contradiction. **Q.E.D.**

If Z is a single point, it is a translate of the trivial abelian subvariety of A. Otherwise we know from the lemmas above that $\dim(\mathrm{Stab}_A(Z)) > 0$, and we can use the induction hypothesis. Let B denote the identity component of $\mathrm{Stab}_A(Z)$. Since $F^m(Z) = Z$, we also have $F^m(B) = B$. Set $\bar{A} := A/B$ and $\bar{Z} := Z/B$, and let $\bar{F}$ denote the endomorphism of $\bar{A}$ induced by F^m. Then we have $\bar{F}(\bar{Z}) = \bar{Z}$, and every eigenvalue of $\bar{F}$ on $\mathrm{Lie}\,\bar{A}$ is an eigenvalue of F^m on $\mathrm{Lie}\,A$ and therefore not a root of unity. By Theorem 2.4, applied to $(\bar{A}, \bar{Z}, \bar{F})$ in place of (A, Y, F), we now deduce that $\bar{Z}$ is a finite union of translates of abelian subvarieties of $\bar{A}$. But Z is irreducible; hence so is $\bar{Z}$. Thus $\bar{Z}$ itself is a translate of an abelian subvariety. That abelian subvariety is equal to A'/B for some abelian subvariety $A' \subset A$ containing B, and so Z is a translate of A', as desired. This finishes the proof of Theorem 2.4, and thus also of Theorems 2.2 and 2.1. **Q.E.D.**

Remark Suppose that in the statement of the preceding theorem, we replace the assumption that none of the eigenvalues of F on $\mathrm{Lie}\,A$ is a root of unity by the weaker assumption that none of the eigenvalues of F on $\mathrm{Lie}\,A$ is an algebraic unit. This case is sufficient for the application to the Manin-Mumford conjecture. Under this weaker assumption, the following alternative proof of Lemma 2.6 can be given; it does not use the theorems of Ueno and Matsumura but only elementary properties of cycle classes.

We work under the hypotheses of Lemma 2.6 and with the notations used above. First, everything can be defined over a countable algebraically closed subfield of K, and this subfield can be embedded into $\mathbb{C}$; thus without loss of generality we may assume that $K = \mathbb{C}$. Then for every integer $i \geq 0$ we abbreviate $H^i := H^i(A, \mathbb{Z})$. Let $c := \mathrm{codim}_A(Z)$ and let $\mathrm{cl}(Z) \in H^{2c}$ be the cycle class of $Z \subset A$. We calculate the cycle class of $(F^m)^{-1}(Z)$ in two ways. On the one hand, we have seen that the group $\mathrm{Ker}(F^m)$ acts transitively on the set of irreducible components of $(F^m)^{-1}(Z)$; the assumption $|G_1| = 1$ implies that it also acts faithfully. Thus the number of irreducible components is $|\,\mathrm{Ker}(F^m)|$. Recall that Z is one of them. Since translation on A does not change cycle classes, we find that all irreducible components of $(F^m)^{-1}(Z)$ have cycle class $\mathrm{cl}(Z)$; hence $\mathrm{cl}\big((F^m)^{-1}(Z)\big) = |\,\mathrm{Ker}(F^m)| \cdot \mathrm{cl}(Z)$. On the other hand F induces a pullback homomorphism $F^* : H^i \to H^i$ for every $i \geq 0$, and by functoriality of cycle classes we have $\mathrm{cl}\big((F^m)^{-1}(Z)\big) = F^*\big(\mathrm{cl}(Z)\big)$. As $\mathrm{cl}(Z)$ is non-zero, we deduce that $|\,\mathrm{Ker}(F^m)|$ is an eigenvalue of F^* on H^{2c}.

Let $d := \dim(A)$, which is also the codimension of any point in A. Repeating the above calculation with the cycle class of a point, we deduce that $|\,\mathrm{Ker}(F^m)|$ is also an eigenvalue of F^* on the highest non-vanishing cohomology group H^{2d}. Since cup product yields isomorphisms $H^i \cong \Lambda^i H^1$ for all $i \geq 0$, which are compatible with F^*, it yields an F^*-equivariant perfect pairing $H^{2c} \times H^{2(d-c)} \to H^{2d}$. From this we deduce that 1 is an eigenvalue of F^* on $H^{2(d-c)}$. Moreover, this eigenvalue must be a product of $2(d-c)$ eigenvalues of F^* on H^1. Now the eigenvalues of F^* on H^1 are precisely those of F on $\mathrm{Lie}\,A$ and their complex conjugates. By assumption, they are algebraic integers but no algebraic units. Thus a product of such numbers can be 1 only if it is the empty product. This shows that $2(d-c) = 0$, that is, $\mathrm{codim}_A(Z) = \dim(A)$; hence Z is a point, as desired. **Q.E.D.**

3. Proof of the Manin-Mumford conjecture

As a preparation let $q = p^r$, where p is a prime number and $r \geq 1$. Let $\mathbb{F}_q$ be the unique field with q elements. Let A be an abelian variety defined over $\mathbb{F}_q$. As in classical algebraic geometry we identify A with the set of its $\overline{\mathbb{F}_q}$-valued points. Let ℓ be a prime number different from p, and let $T_\ell(A)$ be the ℓ-adic Tate module of A, which is a free $\mathbb{Z}_\ell$-module of rank $2\dim(A)$. We denote by φ the Frobenius morphism $A \to A$, which acts on the coordinates of points by taking q-th powers. It also acts on the Tate module via its action on the torsion points. The following result, an analogue of the Riemann hypothesis, is due to Weil (see [14]):

Theorem 3.1 *Let T be an indeterminate. The characteristic polynomial of φ on $T_\ell(A) \otimes_{\mathbb{Z}_\ell} \mathbb{Q}_\ell$ is a monic polynomial in $\mathbb{Z}[T]$. It is independent of ℓ, and all its complex roots have absolute value $\sqrt{q}$.*

Consider now an abelian variety A defined over $\overline{\mathbb{Q}}$ and a closed subvariety X of A. We choose a number field $L \subset \overline{\mathbb{Q}}$ over which both $X \subset A$ can be defined and fix their models over L. For any abelian group G we write $\mathrm{Tor}(G)$ for the group of torsion points of G. Moreover, for any prime p we write $\mathrm{Tor}_p(G)$ for the subgroup of torsion points of order a power of p, and $\mathrm{Tor}^p(G)$ for the subgroup of torsion points of order prime to p. Note that $\mathrm{Tor}(G) = \mathrm{Tor}^p(G) \oplus \mathrm{Tor}_p(G)$.

Choose a prime ideal $\mathfrak{p}$ of $\mathcal{O}_L$ where A has good reduction. Let $\mathbb{F}_q$ be the finite field $\mathcal{O}_L/\mathfrak{p}$, where q is a power of a prime number p. The following lemma is lemma 5.0.10 in [4, p. 105]; we reproduce the proof for the convenience of the reader. We use Weil's theorem 3.1 and reduction modulo $\mathfrak{p}$ to obtain an automorphism of $\overline{\mathbb{Q}}$ and a polynomial that we can feed in Theorem 2.1 to obtain Theorem 1.1.

Lemma 3.2 *There is an element $\sigma_\mathfrak{p} \in \mathrm{Gal}(\overline{\mathbb{Q}}|L)$ and a monic polynomial $P_\mathfrak{p}(T) \in \mathbb{Z}[T]$ all of whose complex roots have absolute value $\sqrt{q}$, such that $P_\mathfrak{p}(\sigma_\mathfrak{p})(x) = 0$ for every $x \in \mathrm{Tor}^p(A)$.*

Proof Let $L_\mathfrak{p}$ be the completion of L at $\mathfrak{p}$. Extend the embedding $L \hookrightarrow L_\mathfrak{p}$ to an embedding of the algebraic closures $\overline{\mathbb{Q}} = \overline{L} \hookrightarrow \overline{L_\mathfrak{p}}$, and the surjection $\mathcal{O}_{L_\mathfrak{p}} \to \mathbb{F}_q$ to a surjection $\mathcal{O}_{\overline{L_\mathfrak{p}}} \to \overline{\mathbb{F}_q}$. On the prime-to-$p$ torsion groups we then obtain natural isomorphisms

$$\mathrm{Tor}^p(A) \cong \mathrm{Tor}^p(A_{\overline{L_\mathfrak{p}}}) \cong \mathrm{Tor}^p(A_{\overline{\mathbb{F}_q}}).$$

The second isomorphism expresses the fact that the field of definition of every prime-to-p torsion point is unramified at $\mathfrak{p}$.

Now, as before, let φ denote the automorphism of $\overline{\mathbb{F}_q}$ and of $\mathrm{Tor}^p(A_{\overline{\mathbb{F}_q}})$ induced by the Frobenius morphism over $\mathbb{F}_q$. Then φ can be lifted to an element $\sigma_\mathfrak{p} \in \mathrm{Gal}(\overline{\mathbb{Q}}|L)$ making the above isomorphisms equivariant. To see this, one first lifts φ to an element $\tau_\mathfrak{p}^{\mathrm{nr}}$ of $\mathrm{Gal}(L_\mathfrak{p}^{\mathrm{nr}}|L_\mathfrak{p})$, where $L_\mathfrak{p}^{\mathrm{nr}}$ is the maximal unramified extension of $L_\mathfrak{p}$. This lifting exists and is unique by [1, Th. 1, p. 26]. This element can then be lifted further to a (non-unique) element $\tau_\mathfrak{p}$ of $\mathrm{Gal}(\overline{L_\mathfrak{p}}|L_\mathfrak{p})$, since $L_\mathfrak{p}^{\mathrm{nr}}$ is a subfield

of $\overline{L_\mathfrak{p}}$. By construction the action of $\tau_\mathfrak{p}$ on $\mathrm{Tor}^p(A_{\overline{L_\mathfrak{p}}})$ corresponds to the action of φ on $\mathrm{Tor}^p(A_{\overline{\mathbb{F}_q}})$. The restriction of $\tau_\mathfrak{p}$ to $\overline{\mathbb{Q}}$ gives the desired element $\sigma_\mathfrak{p}$.

Let $P_\mathfrak{p}(T)$ be the characteristic polynomial of φ on $T_\ell(A_{\overline{\mathbb{F}_q}}) \otimes_{\mathbb{Z}_\ell} \mathbb{Q}_\ell$ for any prime number $\ell \neq p$. By construction, we then have $P_\mathfrak{p}(\varphi) = 0$ on $\mathrm{Tor}_\ell(A_{\overline{\mathbb{F}_q}})$. By Weil's result quoted above, the same equation holds for every prime $\ell \neq p$; so it holds on $\mathrm{Tor}^p(A_{\overline{\mathbb{F}_q}})$. From the construction of $\sigma_\mathfrak{p}$ we deduce that $P_\mathfrak{p}(\sigma_\mathfrak{p}) = 0$ on $\mathrm{Tor}^p(A)$. Finally, Weil's result also describes the complex roots of $P_\mathfrak{p}$. **Q.E.D.**

Let now L^p, $L_p \subset \overline{\mathbb{Q}}$ be the fields generated over L by the coordinates of all points in $\mathrm{Tor}^p(A)$, resp. in $\mathrm{Tor}_p(A)$. Both are infinite Galois extensions of L. Their intersection is known to be finite over L by Serre [10, pp. 33–34, 56–59]. Thus after replacing L by $L^p \cap L_p$, we may assume that L^p and L_p are linearly disjoint over L. The subfield of $\overline{\mathbb{Q}}$ generated by the coordinates of all points in $\mathrm{Tor}(A) = \mathrm{Tor}^p(A) \oplus \mathrm{Tor}_p(A)$ is then canonically isomorphic to $L^p \otimes_L L_p$.

Let $\mathfrak{p}'$ be a second place of good reduction of A, of residue characteristic different from p. Let $\sigma_\mathfrak{p}$, $\sigma_{\mathfrak{p}'}$ and $P_\mathfrak{p}$, $P_{\mathfrak{p}'}$ be the automorphisms and polynomials provided by Lemma 3.2, applied to $\mathfrak{p}$, resp. to $\mathfrak{p}'$. The automorphism of $L^p \otimes_L L_p$ induced by $\sigma_\mathfrak{p} \otimes \sigma_{\mathfrak{p}'}$ then extends to some automorphism σ of $\overline{\mathbb{Q}}$ over L. Since $P_\mathfrak{p}(\sigma_\mathfrak{p})$ vanishes on $\mathrm{Tor}^p(A)$, so does $P_\mathfrak{p}(\sigma)$. Similarly, $P_{\mathfrak{p}'}(\sigma_{\mathfrak{p}'})$ vanishes on $\mathrm{Tor}_p(A) \subset \mathrm{Tor}^{p'}(A)$; hence so does $P_{\mathfrak{p}'}(\sigma)$. Thus with $P(T) := P_\mathfrak{p}(T)P_{\mathfrak{p}'}(T)$ we deduce that $P(\sigma)$ vanishes on $\mathrm{Tor}(A)$. In other words, we have $\mathrm{Tor}(A) \subset \Gamma := \mathrm{Ker}\, P(\sigma)$. With $K = \overline{\mathbb{Q}}$ Theorem 1.1 now follows directly from Theorem 2.1. **Q.E.D.**

References

[1] Algebraic number theory. Proceedings of an instructional conference organized by the London Mathematical Society (a NATO Advanced Study Institute) with the support of the International Mathematical Union. Edited by J. W. S. Cassels and A. Fröhlich. *Academic Press, London; Thompson Book Co., Inc., Washington, D.C.* 1967 xviii+366.

[2] Bogomolov, F. A.: Points of finite order on abelian varieties. (Russian) *Izv. Akad. Nauk SSSR Ser. Mat.* **44** (1980), no. 4, 782–804.

[3] Bogomolov, F. A.: Sur l'algébricité des représentations l-adiques. C. R. Acad. Sci. Paris Sér. A-B **290** (1980), no. 15, A701–A703.

[4] Hrushovski, E.: The Manin-Mumford conjecture and the model theory of difference fields. *Ann. Pure Appl. Logic* **112** (2001), no. 1, 43–115.

[5] Iitaka, S.: Algebraic geometry. An introduction to birational geometry of algebraic varieties. Graduate Texts in Mathematics, 76. North-Holland Mathematical Library, 24. *Springer-Verlag, New York-Berlin*, 1982.

[6] Matsumura, H.: On algebraic groups of birational transformations. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8)* **34** (1963), 151–155.

[7] Pink, R., Roessler, D.: On ψ-invariant subvarieties of semiabelian varieties and the Manin-Mumford conjecture. In preparation.

[8] Raynaud, M.: Courbes sur une variété abélienne et points de torsion. *Invent. Math.* **71** (1983), no. 1, 207–233.
[9] Raynaud, M.: Sous-variétés d'une variété abélienne et points de torsion. *Arithmetic and geometry*, Vol. I, 327–352, Progr. Math. **35**, Birkhäuser Boston, Boston, MA, 1983.
[10] Serre, J.-P.: Oeuvres, vol. IV (1985-1998). *Springer* 2000.
[11] Tate, J. T.: p-divisible groups. *Proc. Conf. Local Fields (Driebergen, 1966)*, 158–183. Springer, Berlin.
[12] Ueno, K.: Classification of algebraic varieties, I. *Compositio Math.* **27** Fasc. 3 (1973), 277–342.
[13] Ullmo, E.: Positivité et discrétion des points algébriques des courbes. *Ann. of Math.* (2) **147** (1998), no. 1, 167–179.
[14] Weil, A.: Variétés abéliennes et courbes algébriques. *Hermann* 1948.
[15] Zhang, S.-W.: Equidistribution of small points on abelian varieties. *Ann. of Math.* (2) **147** (1998), no. 1, 159–165.

ICM 2002 · Vol. I · 547–554

Integrating $\partial\overline{\partial}$

Michael McQuillan*

Abstract

We consider the algebro-geometric consequences of integration by parts.

2000 Mathematics Subject Classification: 32, 14.
Keywords and Phrases: Jensen's formula.

1. Jensen's formula

Recall that for a suitably regular function φ on the unit disc Δ we can apply integration by parts/Stoke's formula twice to obtain for $r < 1$,

$$\int_0^r \frac{dt}{t} \int_{\Delta(t)} dd^c\varphi = \int_{\partial\Delta(r)} \varphi - \varphi(0) \tag{1.1}$$

where $d^c = \frac{1}{4\pi i}(\partial - \overline{\partial})$ so actually we're integrating $\frac{1}{2\pi i}\partial\overline{\partial}$. In the presence of singularities things continue to work. For example suppose $f : \Delta \to X$ is a holomorphic map of complex spaces and $\overline{D}$ a metricised effective Cartier divisor on X, with $f(0) \notin D$, and $\varphi = -\log f^*\|\mathbb{1}_D\|$, where $\mathbb{1}_D \in \mathcal{O}_X(D)$ is the tautological section, then we obtain,

$$\begin{aligned}\int_0^r \frac{dt}{t} \int_{\Delta(t)} f^*c_1(\overline{D}) \;&=\; -\int_{\partial\Delta(r)} \log\|f^*\mathbb{1}_D\| + \log\|f^*\mathbb{1}_D\|(0) \\ &+ \sum_{0<|z|<r} \mathrm{ord}_z(f^*D)\log\left|\frac{r}{z}\right| .\end{aligned} \tag{1.2}$$

Obviously it's not difficult to write down similar formulae for not necessarily effective Cartier divisors, meromorphic functions, drop the condition that $f(0) \notin D$ provided $f(\Delta) \not\subset D$, extend to ramified covers $p : Y \to \Delta$, etc., but in all cases what is clear is,

*Institut des Hautes Etudes Scientifiques, 35 route de Chartres, 91440 Bures-sur-Yvette, France and Department of Mathematics, University of Glasgow G128QW, Scotland. E-mail: mquillan@ihes.fr

Facts 1.3. *(a) If X is compact then $\int_0^r \frac{dt}{t}\int_{\Delta(t)} f^*c_1(\overline{D})$ is very close to being positive if $f(\Delta) \not\subset D$, e.g. in the particular hypothesis preceding (1.2),*

$$\int_0^r \frac{dt}{t}\int_{\Delta(t)} f^*c_1(\overline{D}) \geq \log\|f^*\mathbb{1}_D\|(0) + O_{\overline{D}}(1)$$

(b) There is no such principle for the usual area function $\int_{\Delta(r)} f^*c_1(\overline{D})$ except in extremely special cases such as D ample.

Equally the essence of the study of curves on higher dimensional varieties lies in understanding their intersection with divisors, and, of course, the principle that a curve not lying in a divisor intersects it positively is paramount to the discussion. Consequently the right notion of intersection number for non-compact curves is the so-called characteristic function defined by either side of the identity (1.2). On the other hand intersection number and integration are interchangeable in algebraic geometry, and whence we will write,

Notation 1.4. *Let ω be a $(1,1)$ form on Δ then for $r < 1$,*

$$⨍_{\Delta(r)} \omega := \int_0^r \frac{dt}{t}\int_{\Delta(t)} \omega\,.$$

The more traditional characteristic function notation is reserved for the current associated to a map, i.e.

Definition 1.5. *Let $f : \Delta \to X$ be a map of complex spaces then for $r < 1$, we define,*

$$T_f(r) : A^{1,1}(X) \to \mathbb{C} : \omega \mapsto ⨍_{\Delta(r)} f^*\omega\,.$$

Evidently in many cases one works with forms which are not quite smooth, so there are variations on the definition. In any case in order to motivate our intersection formalism let us pause to consider,

2. Convergence

The basic theorem in the study of subvarieties of a projective variety is Grothendieck's existence and properness of the Hilbert scheme, or if one prefers a sequence of subvarieties of bounded degree has a convergent subsequence. Of course families of smooth curves do not in general limit on smooth curves but rather semi-stable ones, and as such we must necessarily understand convergence of discs in the sense of Gromov [G], i.e.

Definition 2.1. *A disc with bubbles Δ^b is a connected 1-dimensional complex space with singularities at worst nodes exactly one of whose components is a disc Δ and such that every connected component of $\Delta^b\backslash\{\Delta\backslash\mathrm{sing}(\Delta^b)\}$ is a tree of smooth rational curves.*

For $z \in \mathrm{sing}(\Delta^b) \cap \Delta$ and R_z the corresponding tree of rational curves, and provided $0 \notin \mathrm{sing}(\Delta^b)$ we can extend our integral (1.4) to this more general situation by way of,

$$\fint_{\Delta^b(r)} \omega := \fint_{\Delta(r)} \omega + \sum_{z \in \Delta(r) \cap \mathrm{sing}(\Delta^b)} \log \frac{r}{|z|} \int_{R_z} \omega$$

while if $f : \Delta^b \to X$ is a map then we have a graph,

$$\Gamma_f := (\mathrm{id} \times f)(\Delta) \bigcup_{z \in \Delta \cap \mathrm{sing}(\Delta^b)} z \times f(R_z) \subset \Delta \times X \, .$$

An appropriate formulation of Gromov's compactness theorem is then,

Fact 2.2. *Let* $\overline{\mathrm{Hom}}(\Delta, X)$ *be the space of maps from discs with bubbles into a projective variety* X *topologised by way of the Hausdorff metric on the graphs then for* $C : (0,1) \to \mathbb{R}_+$ *any function and* $K \subset \mathrm{Aut}(\Delta)$ *compact the set,*

$$\left\{ f \in \overline{\mathrm{Hom}}(\Delta, X) : \exists \, \alpha \in K \, , \ \fint_{\Delta(r)} \alpha^* f^* c_1(\overline{H}) \leq C(r) \right\}$$

where $\overline{H}$ *is a metricised ample divisor, is compact.*

Under more general hypothesis on X, 2.2 continues to hold, but in the special projective case one has an essentially trivial proof thanks to the ubiquitous Jensen formula, cf. [M] V.3.1. Equally although the appearance of automorphisms looks like an unwarranted complication they are necessitated by,

Remarks 2.3. (a) The possibility of bubbling at the origin.

(b) The defect of positivity for the intersection number as per (1.3)(a).

Observe moreover that the introduction of $\overline{\mathrm{Hom}}(\Delta, X)$ and its precise relation to $\mathrm{Hom}(\Delta, X)$ are both necessary, and easy respectively, i.e.

Fact 2.4. ([M] V.3.5) *Let* $T \supset \mathrm{Hom}(\Delta, X)$ *be such that the bounded subsets in the sense of* 2.2 *are relatively compact then* $T \supset \overline{\mathrm{Hom}}(\Delta, X)$. *Moreover assuming that* X *is not absurdly singular then* $\mathrm{Hom}(\Delta, X) = \overline{\mathrm{Hom}}(\Delta, X)$ *iff* X *contains no rational curves.*

One can equally generalise this to a log, or quasi-projective situation by introducing a divisor D, whose components D_i should be $\mathbb{Q}$-Cartier at which point the appropriate variation thanks to a lemma of Mark Green, [Gr], is,

Fact 2.5. $\overline{\mathrm{Hom}}(\Delta, X \backslash D) \subset \mathrm{Hom}(\Delta, X)$ *iff* $X \backslash D$ *and* $D_i \backslash \bigcup_{j \neq i} D_j$ *do not contain any affine lines.*

In particular $\mathrm{Hom}(\Delta, X \backslash D)$ is relatively compact in $\mathrm{Hom}(\Delta, X)$ if and only if $\overline{\mathrm{Hom}}(\Delta, X \backslash D)$ is compact and the boundary is *mildly hyperbolic* in the sense that $D_i \backslash \bigcup_{j \neq i} D_j$ does not contain affine lines. The latter question is purely algebraic

and closely related to the log minimal model programme. In the case of foliations by curves an even more delicate result holds since as Brunella has observed, [B], the equivalence of $\overline{\mathrm{Hom}}$ with Hom for invariant maps into the orbifold smooth part of a foliated variety is itself equivalent to the said foliated variety being a minimal model.

3. The Bloch principle

Bloch's famous dictum, "Nihil est in infinito quod non fuerit prius in finito", might thus be translated as,

Question 3.1. Suppose for a projective variety X, or more generally a log variety (X, D) there is a Zariski subset Z of $X\backslash D$ through which every non-trivial map $f : \mathbb{C} \to X\backslash D$ must factor then do we have *hyperbolicity modulo* Z, i.e. is it the case that a sequence f_n in $\mathrm{Hom}(\Delta, X)$ not affording a convergent subsequence in $\overline{\mathrm{Hom}}$ must be arbitrarily close (in the compact open sense) to $Z \cup D$.

In the particular case that 2.5 is satisfied we can replace $Z \cup D$ by $\overline{Z}$ and ask for *complete hyperbolicity modulo* Z, but outside of surfaces (2.5) seems difficult to guarantee. Regardless in his thesis Brody, [Br], provided an affirmative answer for both Z and D empty by way of his reparameterisation lemma which was subsequently extended by Green to the case of Z empty and every $D_i \backslash \bigcup_{j \neq i} D_j$ not containing holomorphic lines.

Bearing in mind the singular variant of Green's lemma implicit in 2.5, which for example makes it applicable to stable families of curves, it would appear that the unique known case not covered by the methods of Brody and Green was a theorem of Bloch himself, [Bl], i.e. $\mathbb{P}^2\backslash\{4 \text{ planes in general position}\}$, and its subsequent extension by Cartan to $\mathbb{P}^n$, [C]. However, even here, a moment's inspection shows that 2.5 holds, so one knows a priori that there can be no bubbling, and whence complete hyperbolicity in the sense of 3.1 trivially implies so-called normal convergence modulo the diagonal hyperplanes, and the correct structure is obscured.

Now an extension of the reparameterisation lemma to cover 3.1 would be by far the most preferable way forward, since the non-existence of holomorphic lines is an essentially useless qualitative statement without the quantitative information provided by the convergence of discs. Nevertheless we can vaguely approximate a reparameterisation lemma thanks as ever to Jensen's formula. Specifically consider as given,

Data 3.2.

(a) A $\mathbb{Q}$-Cartier divisor ∂ on a log-variety (X, D).

(b) A sequence $f_n \in \mathrm{Hom}(\Delta, X\backslash D)$ which neither affords a convergent subsequence nor is arbitrarily close to $\partial \cup D$.

In light of (b) we can choose convergent automorphisms $\alpha_n \in \mathrm{Aut}(\Delta)$, such that $\alpha_n^* f_n(0)$ is bounded away from $\partial \cup D$, and given, modulo subsequencing, the convergence of the α_n we may as well suppose this. Moreover for each $0 < r < 1$

we can normalise the current $T_{f_n}(r)$ of 1.5 by its degree with respect to an ample divisor H, which we'll denote by $T^H_{f_n}(r)$ and take a weak limit for a suitable subset $\mathcal{N}$ of $\mathbb{N}$ to obtain a current $T^H_{\mathcal{N}}(r)$. In addition 3.2(b) also tells us that for some fixed $0 < s < 1$, the degrees of the f_n at s go to infinity, and whence by (1.1) and (1.2)

Pre-Fact 3.3. *For $r \geq s$, $T^H_{\mathcal{N}}(r)$ is a positive harmonic current such that, $T^H_{\mathcal{N}}(r) \cdot F \geq 0$ for all effective divisors F supported in $\partial \cup D$.*

What is somewhat less trivial, but once more the key is Jensen's formula, is,

Fact 3.3(bis). ([M] V.2.4) *Subsequencing in $\mathcal{N}$ as necessary, then for $r \geq s$ outside of a set of finite hyperbolic measure (i.e. $(1-r^2)^{-1}dr$) $T^H_{\mathcal{N}}(r)$ is closed.*

Obviously there are various choices involved but whenever we're dealing in the context of countably many projective varieties they can all be rendered functorial, up to a constant, with respect to push forward. The constant itself only causes a problem should it be zero which is usually what one wants to prove anyway, and as such the notation $T(r)$ is relatively unambiguous, and represents in a vague sense a parabolic limit of the sequence f_n.

4. Applications

Applications of course require some knowledge of intersection numbers, and quite generally even for a compact curve $f : C \to X$ there is very little that one can say in general beyond,

Observation 4.1. *Let $f' : C \to P(T_X)$ be the derivative ($\mathbb{P}(\Omega_X)$ in the notation of* EGA*) with L the tautological bundle then,*

$$L_{\cdot f'} C = (2g-2) - \mathrm{Ram}_f \,.$$

This is of course the Riemann-Hurwitz formula if $\dim X = 1$, and there's an equally trivially log-variant where on the right hand side we have to throw in the number of points in the intersection with the boundary D counted without multiplicity the special case of $\mathbb{P}^1 \backslash \{0, 1, \infty\}$ being Mason's "a, b, c" theorem for polynomials. The correct generality for best possible applications is to work with log-smooth Deligne-Mumford stacks (or alternatively just orbifolds since the inertia tends to be irrelevant), however for simplicity let's stick with log-smooth varieties and metricise $T_X(-\log D)$ by way of a complete metric $\| \ \|_{\log}$ on (X, D), which in turn leads to a mildly singular metricisation $\overline{L}$ of the tautological bundle. Supposing for simplicity that $f(0) \notin D$ with f unramified at the origin then Jensen's formula yields,

Observation 4.2(bis). *Notations as above,*

$$\begin{aligned} \fint_{\Delta(r)} f^*c_1(\overline{L}) &= -\log\left\|f_*\left(\frac{\partial}{\partial z}\right)\right\|_{\log}(0) + \int_{\partial\Delta(r)} \log\left\|f_*\left(\frac{\partial}{\partial z}\right)\right\|_{\log} \\ &+ \sum_{0<|z|<r} \min\{1, \mathrm{ord}_z(f^*D)\}\log\frac{r}{|z|} \\ &- \sum_{\substack{0<|z|<r \\ f(z)\notin D}} \mathrm{ord}_z(R_f)\log\frac{r}{|z|}\,. \end{aligned}$$

Combining the concavity of the logarithm and once more Jensen's formula, but this time for $dd^c \log\log^2 \|\mathbb{1}_D\|$ for any norm on the boundary divisor D, immediately yields in the notations of 3.3,

Fact 4.3. *Let $T'(r)$ be the current associated to the logarithmic derivative of a sequence $f_n \in \mathrm{Hom}(\Delta, X\backslash D)$ with $f_n(0)$ not arbitrarily close to D and which does not afford a convergent subsequence then outwith a set of finite hyperbolic measure,*

$$L.\,T'(r) \leq 0\,.$$

The so-called tautological inequality 4.3 is well adapted for applications to convergence of discs (note incidentally that it's implicit to the formulation that a smooth metric on the bundle $T_X(-\log D)$ is being employed). Nevertheless for more delicate questions such as quantifying degenerate/non-convergent behaviour. etc. there is a wealth of information in (4.2) that is lost in the coarser corollary. Indeed even using the concavity of the logarithm distorts a very delicate term measuring the 'ramification at ∞', i.e. the distorsion of the boundary from it's length in the Poincaré metric, which is closely related to the difficulty of extracting an isoperimetric inequality from a knowledge of hyperbolicity in the sense of 3.1. While from the still deeper curvature point of view, 4.2 is simply a doubly integrated tautological Schwarz lemma, since by definition metricising $T_X(-\log D)$ by way of a metric ω of curvature $\leq -K$ is equivalent to a lower bound of the left hand side of the form,

$$K \fint_{\Delta(r)} \omega$$

for all infinitely small, and whence all in the large, possible discs. While on the subject of curvature and isoperimetric inequalities, a variant specific to dimension 1 replaces the current δ_D implicitly hidden in (4.2) by the current associated to the boundary of a simply connected region, i.e.

Variant 4.4. *Suppose* $\dim X = 1$, *let $U_i \subset X$ be simply connected, $h_i : \Delta \xrightarrow{\sim} U_i$ isomorphisms and put*

$$\delta_{\Gamma_i} : A^0(X) \to \mathbb{C} : \varphi \mapsto \int_{\partial\Delta} h_i^*\varphi\,.$$

Then specific to dimension 1, δ_{Γ_i} is closed and may be written, $c_1(H) + dd^c\gamma_i$ for H an ample divisor of degree 1. Now apply Jensen's formula to recover an integrated form of Ahlfor's isoperimetric inequality, and the Five Island's Theorem.

Returning to varieties and divisors it's still possible to employ (4.2) to get integrated isoperimetric inequalities for more general situations that preserve some 1-dimensional flavour, i.e. discs which are invariant by foliations by curves, with *canonical foliation singularities* (with the obvious definition of that notion which is functorial with respect to the ideas) and which do not pass through the singularities. The latter hypothesis which is reasonable for the study of the leafwise variation of the Poincaré metric is however somewhat restrictive for other applications, and is probably unnecessary as suggested by the essentially optimal inequality of [M] V.4.4 for foliations on surfaces which employs (4.2) to a very large number of monoidal transformations in the foliation singularities. Regardless here is a genuinely 2-dimensional theorem,

Theorem 4.5. ([M] V.5) *Let (X, D) be a smooth logarithmic surface with $\Omega_X(\log D)$ big (e.g. log-general type, and $s_2(\Omega_X(\log D)) > 0$) then there is a proper Zariski subset Z of $X\backslash D$ such that $X\backslash D$ is complete hyperbolic (in the sense of* 3.1 *et sequel) modulo Z.*

Indeed one can even optimally quantify (cf. op. cit.) the degeneration of the Kobayashi metric (which is evidently continuous and non-zero off Z) around Z. Amusingly the theorem only covers $\mathbb{P}^2\backslash\{5$ planes in general position$\}$, although it's a good exercise in the techniques (cf. op. cit. V.4) to prove Bloch's theorem too, at which point a rather small sequence of blow ups replaces all of the original estimation. In any case (4.5) should only be seen as a stepping stone which in order of ascending difficulty leaves open the following questions, viz,

Concluding Remarks 4.6. For concreteness take a smooth algebraic surface X of general type with $c_1^2 > c_2$ (otherwise the following should be understood in terms of higher jets, but not for anything more general than a surface) then,

(a) Do we have an isoperimetric inequality with appropriate degeneration along the subset Z of (4.5).
(b) Is the Kobayashi metric negatively curved.
(c) For each $x \notin Z$ and t a tangent direction at x, is there a unique up to the usual action of $\mathrm{SL}_2(\mathbb{R})$ pointed disc with maximal tangent in the direction t, and if so does it continue to be so along its image, i.e. is there a continuous (off Z) connection whose geodesics are the discs defining the Kobayashi metric.

5. Thanks

In closing it is a pleasure to thank M. Brunella for introducing me to the unit disc and bubbling, together with M. Gromov for continuing this education, but above all Cécile without whom the reader could never have got this far.

References

[Bl] Bloch A., Sur les systèmes de fonctions holomorphes à variétés linéaires lacunaires, *Ann. École Norm. Sup.* **43** (1926), 309–362.

[Br] Brody R., Compact manifolds and hyperbolicity, *Trans. Am. Math. Soc.* **235** (1978), 213–219.

[B] Brunella M., A subharmonic variation of the leafwise Poincaré metric, *Univ. Dijon Preprint*, 2001.

[C] Cartan, H., Sur les systèmes de fonctions holomorphes à variétés linéaires lacunaires et leurs applications, *Ann. École Norm. Sup.* **45** (1928), 255–346.

[G] Gromov, M. Pseudo-holomorphic curves in symplectic manifolds, *Invent. Math.* **82** (1985), 307–347.

[Gr] Green, M. The hyperbolicity of the complement of $(2n + 1)$ hyperplanes in general position in $\mathbb{P}^n$, and related results, *Proc. Amer. Math. Soc.* **66** (1977), 109–113.

[M] McQuillan M., Canonical models of foliations, cf. IHES preprints IHES/M/01/42 and IHES/M/01/59.

ICM 2002 · Vol. I · 555–556

Hidden Markov and State Space Models Asymptotic Analysis of Exact and Approximate Methods for Prediction, Filtering, Smoothing and Statistical Inference

P. Bickel* Y. Ritov† T. Ryden‡

Abstract

State space and hidden Markov models can both be subsumed under the same mathematical structure. On a suitable probability space $(\Omega, \mathcal{A}, P)$ are defined $(X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n, \ldots)$ a sequence of random "variables" taking values in a product space $\prod_{j=1}^{\infty}(\mathcal{X}_j \times \mathcal{Y}_j)$ with an appropriate sigma field. The joint behavior under P is that the X_j are stationary Markovian and that given $(X_1, X_2, \ldots)$ the Y_j are independent and further that Y_j is independent of all $X_i : i \neq j$ given X_j. If $\mathcal{H}$ is finite these are referred to as Hidden Markov models. The general case though focussing on $\mathcal{X}$ Euclidean is referred to as state space models. Essentially we observe only the Y's and want to infer statistical properties of the X's given the Y's. The fundamental problems of filtering, smoothing prediction are to give algorithms for computing exactly or approximately the conditional distribution of X_t given $(Y_1, \ldots, Y_t)$ (Filtering), the conditional distribution of X_t given $Y_1, \ldots, Y_T$, $T > t$ (Smoothing) and the conditional distribution of $X_{t+1}, \ldots, X_T$ given $Y_1, \ldots, Y_t$ (Prediction). If as is usually the case P is unknown and is assumed to belong to a smooth parametric family of probabilities $\{P_\theta : \theta \in R^d\}$,we face the further problem of efficiently estimating θ using $Y_1, \ldots, Y_T$ (computation of the likelihood, and maximum likelihood estimation, etc.).

State space models have long played an important role in signal processing. The Gaussian case can be treated algorithmically using the famous Kalman filter [6]. Similarly since the 1970s there has been extensive application of Hidden Markov models in speech recognition with prediction being the most important goal. The basic theoretical work here, in the case $\mathcal{X}$ and $\mathcal{Y}$ finite (small) providing both algorithms and asymptotic analysis for inference is that of Baum and colleagues [1]. During the last 30-40 years these general models have proved of great value in applications ranging from genomics to finance—see for example [7].

*University of California, Berkeley, USA. E-mail: bickel@stat.berkeley.edu
†Hebrew University, Israel
‡University of Lund, Sweden

Unless the X, Y are jointly Gaussian or $\mathcal{X}$ is finite and small the problem of calculating the distributions discussed and the likelihood exactly are numerically intractable and if $\mathcal{Y}$ is not finite asymptotic analysis becomes much more difficult. Some new developments have been the construction of so-called "particle filters" (Monte Carlo type) methods for approximate calculation of these distributions (see Doucet et al. [4]) for instance and general asymptotic methods for analysis of statistical methods in HMM [2] and other authors.

We will discuss these methods and results in the light of exponential mixing properties of the conditional (posterior) distribution of $(X_1, X_2, \dots)$ given $(Y_1, Y_2, \dots)$ already noted by Baum and Petrie [1] and recent work of the authors Bickel, Ritov and Ryden [3], Del Moral and Jacod in [4], Douc and Matias [5].

2000 Mathematics Subject Classification: 60, 62.

References

[1] L. E. Baum & T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, *Ann. Math. Statist.*, 37 (1966), 1554–1563.

[2] P. Bickel, Y. Ritov & T. Ryden, Asymptotic normality of the maximum likelihood estimate for HMM, *Ann. Statist.*, (1998), 1614–1635.

[3] P. Bickel, Y. Ritov & T. Ryden,HMM Likelihoods behave like IID ones,to appear in *Annales de l'Institut Henri Poincare (Probabilites)*

[4] A.Doucet,N.-de Freitas,N.-Gordon,eds. *Sequential Monte Carlo in Practice*, Springer, 2001.

[5] R. Douc & C. Matias, Asymptotics of the maximum likelihood estimator for general hidden Markov models, *Bernoulli*, 7 (2001), 631–699.

[6] R. E. Kalman, A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, 82 (1960), 35–45.

[7] I. L. MacDonald & W. Zucchini, *Hidden Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall, London, 1997.

ICM 2002 · Vol. I · 557–566

Statistical Equivalence and Stochastic Process Limit Theorems*

Lawrence D. Brown†

Abstract

A classical limit theorem of stochastic process theory concerns the sample cumulative distribution function (CDF) from independent random variables. If the variables are uniformly distributed then these centered CDFs converge in a suitable sense to the sample paths of a Brownian Bridge. The so-called Hungarian construction of Komlos, Major and Tusnady provides a strong form of this result. In this construction the CDFs and the Brownian Bridge sample paths are coupled through an appropriate representation of each on the same measurable space, and the convergence is uniform at a suitable rate.

Within the last decade several asymptotic statistical-equivalence theorems for nonparametric problems have been proven, beginning with Brown and Low (1996) and Nussbaum (1996). The approach here to statistical-equivalence is firmly rooted within the asymptotic statistical theory created by L. Le Cam but in some respects goes beyond earlier results.

This talk demonstrates the analogy between these results and those from the coupling method for proving stochastic process limit theorems. These two classes of theorems possess a strong inter-relationship, and technical methods from each domain can profitably be employed in the other. Results in a recent paper by Carter, Low, Zhang and myself will be described from this perspective.

1. Probability setting

1.1. Background

Let F be the CDF for a probability on [0,1];. F abs. cont., with

$$f(x) \triangleq \frac{\partial F}{\partial x} \text{ on } [0,1].$$

*Research supported in part by NSF-Division of Mathematical Sciences.

†Statistics Department, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, USA. E-mail: lbrown@wharton.upenn.edu

Let $X_1, \ldots, X_n$ iid from F. $\hat{F}_n$ denotes the sample CDF,

$$\hat{F}_n(x) \triangleq \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}_{[0,x]}(X_j).$$

Let $\hat{Z}_n$denote the corresponding sample "bridge",

$$\hat{Z}_n(x) \triangleq \hat{F}_n(x) - F(x) \tag{1}$$

Let W(t) denote the standard Wiener process on [0,1] and let $\hat{W}_n$ denote the white noise process with drift f and local variance $f(t)/n$. Thus $\hat{W}_n$ solves

$$d\hat{W}_n(t) = f(t)dt + \sqrt{\frac{f(t)}{n}}dW(t).$$

An alternate description of $\hat{W}_n$ is that it is the Gaussian process with mean F(t) and independent increments having

$$var\left(\hat{W}_n(t) - \hat{W}_n(s)\right) = \frac{1}{n}\left(F(t) - F(s)\right), \ for\ 0 \le s < t \le 1.$$

The analog of $\hat{Z}_n$ is the Gaussian Bridge, defined by

$$\hat{B}_n(t) = \frac{\hat{W}_n(t)}{\hat{W}_n(1)} - F(t).$$

There are various ways of describing the stochastic similarity between $\hat{Z}_n$ and $\hat{B}_n$. For example Komlos, Major, and Tusnady (1975, 1976) proved a result of the form

Theorem (KMT): *Given any absolutely continuous F* $\{X_1,\ldots,X_n\}$ *can be defined on a probability space on which* $\hat{B}_n$ *can also be defined as a (randomized) function of* $\{X_1,\ldots,X_n\}$. *This can be done in such a way that* $\hat{B}_n$ *has the Gaussian Bridge distribution, above, and*

$$P_F\left(\sup_{t\in[0,1]} \sqrt{n}\left|\hat{Z}_n(t) - \hat{B}_n(t)\right| > a_n\right) \le c. \tag{2}$$

Here $c > 0$ *and* a_n *are suitable positive constants with* $a_n \sim (d\log n)/\sqrt{n}$ *for some* $d > 0$. *The process* $\hat{B}_n$ *can be constructed as a (randomized) function of* $\hat{Z}_n$, *that is,* $\hat{B}_n(t) = Q_n\left(\hat{Z}_n(t)\right)$. *It should be noted that the construction depends on knowledge of F.*

[Various authors, such as Csörgö and Revesz (1981) and Bretagnolle and Massart (1989) have given increasingly detailed and precise values for a_n and $c = c(a_n)$, and also uniform (in n) versions of (2). These are not our focus.]

1.2. Extensions

1. Results like the above also extend to functional versions of the process $\hat{Z}_n$. Various authors including Dudley (1978), Massart (1989) and Koltchinskii (1994) have established results of the following form.

Let q:[0,1]→ $\Re$ be of bounded variation. One can define

$$\hat{Z}_n(q) \triangleq \int q d\{\hat{F}_n - F\} = \int \left(F - \hat{F}_n\right) dq.$$

(Thus, $\hat{Z}_n(x) = \hat{Z}_n\left(\mathbf{I}_{[0,x]}\right)$.) There is a similar definition for $\hat{B}_n(q)$ as a stochastic integral. (See, for example, Steele (2000).) Then the KMT theorem extends to a fairly broad, but not universal, class of functions, $\mathcal{Q}$. That is, for each F, $\hat{B}_n$ can be defined to satisfy

$$P_F\left(\sup_{q\in Q}\sqrt{n}\left|\hat{Z}_n(q) - \tilde{B}_n(q)\right| > a'_n\right) \leq c \text{ where } \{a'_n\} \text{ depends on } \mathcal{Q}. \tag{3}$$

(For most classes $\mathcal{Q}$, $a'_n\sqrt{n}/\log n \to \infty$ so that $a'_n >> a_n$.)

2. Bretagnolle and Massart (1989) proved a similar result for inhomogeneous Poisson processes. Let $\{T_1,\ldots,T_N\}$ be (ordered) observations from an inhomogeneous Poisson process with cumulative intensity function nF and, correpondingly, (local) intensity nf. Note that N∼Poisson(n) and conditionally given N the values of $\{T_1,\ldots,T_N\}$ are the order statistics corresponding to an iid sample from the distribution F. In this context we continue to define $\hat{F}_n(t) = n^{-1}\left\{\sum_{j=1}^{N} I_{[0,t]}(T_j)\right\}$ where the term in braces now has a Poisson distribution with mean nF(t). Also, continue to define $\hat{Z}_n(t) \triangleq \hat{F}_n(t) - F(t)$ as in (1). (But, note that it is no longer true that $\hat{Z}_n(1) = 0$, w.p.1, as was the case in (1).)

Then versions of the conclusions (2) and (3) remain valid. We give an explicit statement since this result will provide a model for our later development.

Theorem (BM): *Given any n and any absolutely continuous F the observations* $\{T_1,\ldots,T_N\}$ *of the inhomogeneous Poisson process can be defined on a probability space on which* $\hat{B}_n$ *can also be defined as a (randomized) function of* $\{T_1,\ldots,T_N\}$. *This can be done in such a way that* $\hat{B}_n$ *has the Gaussian Bridge distribution, above, and*

$$P_F\left(\sup_{t\in[0,1]}\sqrt{n}\left|\hat{Z}_n(t) - \hat{B}_n(t)\right| > a_n\right) \leq c. \tag{4}$$

Here $c > 0$ *and* a_n *are suitable constants with* $a_n \sim d\log n/\sqrt{n}$.

Remark: Clearly there must be extensions of (3) that are valid for the Poisson case also, although we are not aware of an explicit treatment in the literature. Such a statement would conclude in this setting that

$$P_F\left(\sup_{q\in Q}\sqrt{n}\left|\hat{Z}_n(q) - \hat{B}_n(q)\right| > a'_n\right) \leq c \text{ where } \{a'_n\} \text{ depends on } \mathcal{Q}. \tag{5}$$

2. Main results

The objective is a considerably modified version of (3) and (5) that is stronger in several respects and (necessarily) different in others. We will concentrate for most of the following on the statement (5) since our results are slightly stronger and more natural in this setting. The extension of (3) will be deferred to a concluding Section.

Expression (5) involves the target function $\hat{B}_n$. In the modified version the role of target function is instead played by $\tilde{W}_n$ which is the solution to the stochastic differential equation

$$d\tilde{W}_n(t) = g(t)dt + \frac{1}{2\sqrt{n}}dW(t) \tag{6}$$

where $g(t) = \sqrt{f(t)}$. An alternate description of $\tilde{W}_n$ is thus

$$\tilde{W}_n = G(t) + W(t)/(2\sqrt{n}) \text{ where } G(t) = \int_0^t \sqrt{f(\tau)}d\tau. \tag{7}$$

(In the special case where f is the uniform density, f=1, then $\tilde{W}_n = W_{4n}$.)

The role of the constructed random process $\hat{Z}_n$ is now played by a differently constructed process $\tilde{Z}_n$. As before $\tilde{Z}_n$ depends only on $\{T_1,...T_N\}$, and not otherwise on their CDF, F. This version also involves a large set, $\mathcal{F}$, of absolutely continuous CDFs. Both $\tilde{Z}_n$ and $\mathcal{F}$ will be described later in more detail. Here are statements of the main results.

Theorem 1: *Let $\mathcal{F}$ be a set of densities satisfying Assumption A or A', below. Let $\mathcal{Q}$ be the set of all functions of bounded variation. Let $\{T_1,\ldots,T_N\}$ be an inhomogeneous Poisson process with local intensity nf. The process $\tilde{Z}_n$ can be constructed as a (randomized) function of $\{T_1,\ldots,T_N\}$, with the construction not depending on f. The Gaussian process $\tilde{W}_n$ having the distribution (7) can also be defined on this same space as a (randomized) function of $\{T_1,\ldots,T_N\}$. [This construction depends on f on a set of probability at most c_n.] This can be done in such a way that*

$$\sup_{f\in F} P_f\left(\sup_{q\in Q}\left|\tilde{Z}_n(q) - \tilde{W}_n(q)\right| > 0\right) \le c_n \to 0. \tag{8}$$

To be more precise, the phrase in brackets refers to the fact that there is a basic construction, independent of f, and that this construction must then be modified on a set of measure at most c_n with this set and the modification depending on f.

For the situation of iid variables, as in (1), a similar result holds. In this case the matching Gaussian process is again $\tilde{W}_n$, rather than the Brownian bridge of the KMT theorem.

Theorem 2: *Let $\mathcal{F}$ be a set of densities satisfying Assumption B, below. Let $\mathcal{Q}$ be the set of all functions of bounded variation. Given any n and f$\in\mathcal{F}$, iid variables $\{X_1,\ldots,X_n\}$ with density f can be defined on a probability space. A process $\tilde{\tilde{Z}}_n$ can*

be constructed as a (randomized) function of $\{X_1,\ldots,X_n\}$*, with the construction not depending on* f*. The Gaussian process* $\tilde{W}_n$ *having the distribution (7) can also be defined on this same space as a (randomized) function of* $\{X_1,\ldots,X_n\}$*. [This construction depends on* f*, but only on a set of probability at most* c_n*.] This can be done in such a way that*

$$\sup_{f\in\mathcal{F}} P_f\left(\sup_{q\in Q}\left|\tilde{\tilde{Z}}_n(q)-\tilde{W}_n(q)\right|>0\right)\le c_n\to 0. \tag{9}$$

3. Statistical background

3.1. Settings

The first purpose of the discussion here is to motivate the probabilistic results described above. A second purpose is to state the result on which to base the proof of Theorem 1. The setting involves two statistical formulations:

Formulation 1 (nonparametric inhomogeneous Poisson process): The observations are $\mathbf{T}=\{\mathrm{T}_1,\ldots,\mathrm{T}_N\}$ from the Poisson process with local intensity nf, $f\in\mathcal{F}$. The problem is "nonparametric" because the "parameter space", $\mathcal{F}$, is a very large set – too large to be smoothly parameterized by a mapping from a (subset of) a finite dimensional Euclidean space. Some possible forms for $\mathcal{F}$ are discussed below. The statistician desires to make some sort of inference, δ, (possibly randomized) based on the observation of $\mathbf{X}$.

Formulation 1' (nonparametric density with random sample size): The relation between Poisson processes and density problems has been mentioned above. As a consequence, Problem 1 is equivalent to a situation where the observations are $\{\mathrm{X}_1,\ldots,\mathrm{X}_N\}$ with N~Poisson(n) and $\{\mathrm{X}_1,\ldots,\mathrm{X}_N\}$ the order statistics from a sample of size N from the distribution with density f. Clearly, this situation is closely related to the more familiar one in which the observations are $\{\mathrm{X}_1,\ldots,\mathrm{X}_n\}$ with n specified in advance.

Formulation 1" (nonparametric density with fixed sample size): This formulation refers to the more conventional density setting in which the observations are $\{\mathrm{X}_1,\ldots,\mathrm{X}_n\}$ iid with density f.

Formulation 2 (white noise with drift): The statistician observes a White noise process $d\tilde{W}_n(t)$, $t\in[0,1]$, with drift $g\in\mathcal{G}$ and local variance $1/4n$. Thus

$$d\tilde{W}_n(t)=g(t)dt+\frac{1}{2\sqrt{n}}dW(t),$$

and $\tilde{W}_n(t)-G(t)=W(t)/2\sqrt{n}$ where $G(t)=\int_0^t g(\tau)d\tau$. Again $\mathcal{G}$ is a very large – hence "nonparametric" – parameter space. Throughout, $\mathcal{G}\subset\mathcal{L}_2=\{g:\int g^2<\infty\}$. As of now, there need be no relation between f in Formulation 1 and g in Formulation 2, but such a relation will later be assumed in connection with Theorem 1, where

$$g=\sqrt{f}\ \text{ and }\ \mathcal{G}=\left\{\sqrt{f}:f\in\mathcal{F}\right\}. \tag{10}$$

This can alternatively be considered as a statistical formulation having parameter space $\mathcal{F}$ under the identification (10). We take this point of view in the BCLZ theorem, below.

3.2. Constructive asymptotic statistical equivalence

Here is one definition of the strongest form of such an equivalence.

Definition (asymptotic equivalence): *Let* $\mathcal{P}_j^{(n)} = (\mathcal{X}_j^{(n)}, \mathcal{B}_j^{(n)}, \mathcal{F}_j^{(n)})$ $j = 1,2$, $n = 1,2,\ldots$ *be two sequences of statistical problems on the same sequence of parameter spaces,* $\Theta^{(n)}$. *Hence,* $\mathcal{F}_j^{(n)} = \left\{F_{j,\theta}^{(n)} : \theta \in \Theta^{(n)}\right\}$. *Then* Π_1 *and* Π_2 *are asymptotically equivalent if there exist (randomized) mappings* $Q_j^{(n)} : \mathcal{X}_j^{(n)} \to \mathcal{X}_k^{(n)}$, $j, k = 1, 2, k \neq j$, *such that*

$$\sup_{\theta \in \Theta^{(n)}} \left\| F_{j,\theta}^{(n)}(\cdot) - \int Q_k^{(n)}(\cdot|x_k)\, F(dx_k) \right\|_{TV} = c_n \to 0, j, k = 1, 2, k \neq j, \qquad (11)$$

where $\|\cdot\|_{TV}$ *denotes the total variation norm.*

This definition involves a reformulation of the general theory originated by LeCam (1953, 1964). See also Le Cam (1986), Le Cam and Yang (2000), van der Vaart (2002) and Brown and Low (1996) for background on this theory including several alternate versions of the definition and related concepts, a number of conditions that imply asymptotic equivalence, and many applications to a variety of statistical settings. Note that both Formulations 1 and 2 involve an index, n, and can thus be considered as sequences of statistical problems in the sense of the definition.

3.3. Spaces of densities (or intensities)

Suitable families of densities, $\mathcal{F}$, can be defined via Besov norms with respect to the Haar basis. The Besov norm with index α and shape parameters p = q can most conveniently be defined via the stepwise approximants to f at resolution level k. These approximants are defined as

$$\bar{f}_k(t) = \sum_{\ell=0}^{2^k-1} I_{[\ell/2^k,(\ell+1)/2^k)}(t) \int_{\ell/2^k}^{(\ell+1)/2^k} 2^k f,$$

and the Besov(α,p) norm is defined as

$$\|f\|_{\alpha,p} = \left\{ \left|\bar{f}_0\right|^p + \sum_{k=0}^{\infty} 2^{pk\alpha} \left\|\bar{f}_k - \bar{f}_{k+1}\right\|_p^p \right\}^{1/p}.$$

The statement of Theorem 1 can now be completed by stating the assumption on $\mathcal{F}$ needed for its validity.

Assumption A: *$\mathcal{F}$ satisfies*

$$\mathcal{F} \subset \left\{ f : \inf_{0 \le x \le 1} f(x) \ge \varepsilon_0 \right\} \text{ for some } \varepsilon_0 > 0 \tag{12}$$

and $\mathcal{F}$ is compact in both Besov(1/2,2) and Besov(1/2,4).

Other function spaces are also conventional for nonparametric statistical applications of this type. The most common of these are based on either the Lipshitz norm $\|f\|_\beta^{(L)}$ or the Sobolev norm $\|f\|_\beta^{(S)}$. These are defined for $\beta \le 1$ by

$$\|f\|_\beta^{(L)} = \sup_{0 \le x < y \le 1} \frac{|f(y) - f(x)|}{|y - x|^\beta}, \quad \|f\|_\beta^{(S)} = \sum_{-\infty}^{\infty} k^{2\beta} \vartheta_k^2$$

where $\vartheta_k = \int_0^1 f(x) e^{ik2\pi x} dx$ denote the Fourier coefficients of f. (Both spaces have natural definitions for β¿1 as well, but we need consider here only the case $\beta \le 1$.)

The following implies Assumption A and hence also suffices for validity of Theorem 1.

Assumption A': *$\mathcal{F}$ satisfies (12), and is bounded in the Lipshitz norm with index β, and is compact in the Sobolev norm with index α, where $\alpha \ge \beta$ and either $\beta > 1/2$ or $\alpha \ge 3/4$ and $\alpha + \beta \ge 1$.*

The following assumption is noticeably stronger than either A' or A, and is used in Theorem 2.

Assumption B: *$\mathcal{F}$ satisfies (12) and is bounded in the Lipshitz norm with index β, where $\beta > 1/2$.*

For more information about the relation of these spaces in this context see Brown, Cai, Low and Zhang (2002) and Brown, Carter, Low and Zhang (2002) (referred to as BCLZ below).

3.4. Statistical equivalence theorems

BCLZ then extended earlier results of Nussbaum (1996) and Klemela and Nussbaum (1998) to prove the following basic result:

Theorem a (BCLZ): *Consider the statistical Formulations 1 and 2 with the parameter space $\mathcal{F}$ and the relation (10). Assume $\mathcal{F}$ satisfies Assumption A (or A'. Then the sequences of statistical problems defined in these two formulations are asymptotically statistically equivalent.*

BCLZ describes in detail a construction of $\tilde{Z}_n$ as a (randomized) function of $\{T_1, \ldots, T_n\}$. (More precisely, BCLZ describes the construction of the Haar basis representation of $\tilde{Z}_n$, from which $\tilde{Z}_n$ can directly be recovered.) This construction is invertible, in that $\{T_1, \ldots, T_n\}$ can be recovered as a function of $\tilde{Z}_n$. Further, BCLZ shows that both $\tilde{Z}_n$and $\tilde{W}_n$ can be represented on the same probability space so that their distributions, $P_{\tilde{Z}_n}$ and $P_{\tilde{W}_n}$, say, satisfy

$$\left\| P_{\tilde{Z}_n} - P_{\tilde{W}_n} \right\|_{TV} \to 0.$$

The mappings $\{Q_j^{(n)}$: j=1,2, n = 1,2,...$\}$ that yield the equivalence of the above theorem can then be directly inferred from this construction. To save space here we refer the reader to that paper or Brown (2002) for details of the construction and proof. It can be remarked that these bear considerable similarity to parts of the construction and proof in Bretagnolle and Massart (1989) and other proofs of KMT type theorems. But there are also some basic differences, especially those related to the appearance of the square-root in the fundamental relation (10) and the total variation norm in the definition of equivalence. In addition, the fact that (8) is uniform in $\mathcal{Q}$ and $\mathcal{F}$ entails the need for various refinements in the proof.

Theorem 1 is now an immediate logical consequence of this result from BCLZ and the following lemma.

Lemma: *Suppose* $\mathcal{P}_j^{(n)} = (\mathcal{X}_j^{(n)}, \mathcal{B}_j^{(n)}, \mathcal{F}_j^{(n)})$ *j = 1,2, n = 1,2,... are asymptotically equivalent sequences of statistical problems on the same sequence of parameter spaces,* $\Theta^{(n)}$. *Let* $\{Q_j^{(n)}$: *j=1,2, n=1,2,...*$\}$ *denote a sequence of mappings that define this equivalence, as in (11). Then there are non-randomized mappings* $\{\tilde{Q}_j^{(n)}$: *j=1,2, n = 1,2,...*$\}$ *such that*

$$P_f(\tilde{Q}_j^{(n)} = Q_j^{(n)}) \geq 1 - c_n \text{ for every } f \in \mathcal{F}_j^{(n)},\ j = 1,2, n = 1,2,... \tag{13}$$

and for every $\theta \in \Theta^{(n)}$

$$P_{f_{j,\theta}}\left(\tilde{Q}_j^{(n)}\left(X_j^{(n)}\right) \in A\right) = P_{f_{k,\theta}}\left(X_k^{(n)} \in A\right), \theta \in \Theta^{(n)} \tag{14}$$

for every measurable $A \subset \mathcal{X}_k^{(n)}$, $j,k = 1,2,\ j \neq k,\ n = 1,2,\cdots$.

Proof of Lemma: Fix n, j, k≠j, $\theta \in \Theta^{(n)}$. Let F_k denote the distribution under θ of $X_k^{(n)}$ and let F_k' denote the distribution under θ of $Q_j^{(n)}\left(X_j^{(n)}\right)$. Let H = min($F_k$, F_k'). Let $\infty \geq f_k' = \frac{dF_k'}{dH} \geq 1$. Then define $\tilde{Q}_j^{(n)}$ as a version of the randomized map satisfying

$$\tilde{Q}_j(B|x) = \frac{1}{f_k'(x)} Q_j(B|x) + \frac{f_k' - 1}{f_k'}\left(F_k'(B) - H(B)\right).$$

This completes the proof of the lemma, and consequently also that of Theorem 1. □

Theorem 2 requires a slightly different fundamental result. The following result is the foundation for the proof of Theorem 2. It is adapted from Theorem 2 of BCLZ. This result closely resembles Theorem a, above, but as noted in BCLZ it appears to require a modified construction for its proof. The argument there is based heavily on results in Carter (2001).

Theorem b (BCLZ): *Consider the statistical Formulations 1" and 2 with the parameter space* $\mathcal{F}$ *and the relation (10). Assume* $\mathcal{F}$ *satisfies Assumption B. Then the sequences of statistical problems defined in these two formulations are asymptotically statistically equivalent.*

References

[1] Bretagnolle, J and Massart, P. (1989) Hungarian constructions from the nonasymptotic viewpoint, *Ann. Probab., 17*, 239–256.

[2] Brown, L. D. (2002) The analogy between statistical equivalence and stochastic strong limit theorems. Preprint available via http://ljsavage.wharton.upenn.edu/~lbrown/

[3] Brown, L. D., Cai, T. T., Low, M. G. and Zhang, C-H. (2002a) Asymptotic equivalence theory for nonparametric regression with random design, *Ann. Statist.* 30, 688–707.

[4] Brown, L. D., Carter, A., Low, M. G. and Zhang, C-H. (2002b) (BCLZ) Asymptotic equivalence theory for a Poisson process with variable intensity. Preprint, available via http://ljsavage.wharton.upenn.edu/~lbrown/

[5] Carter, A. (2001) Deficiency distance between multinomial and multivariate normal experiments under smoothness constraints on the parameter set, Preprint available via http://www.pstat.ucsb.edu/faculty/carter/research.html

[6] Csörgö, M and Revesz, P. (1981) *Strong Approximations in Probability and Statistics*, Academic Press, NY.

[7] Dudley, R. M. (1978) Central limit theorems for empirical measures, *Ann. Probab. 6*, 899–929.

[8] Klemela, J. and Nussbaum, M. (1998) Constructive asymptotic equivalence of density estimation and Gaussian white noise, Preprint available at http//www.math.cornell.edu/~nussbaum.

[9] Koltchinskii, V. I. (1994) Komlos-Major-Tusnady approximation for the general empirical process and Haar expansions of classes of functions, *Jrnl of Theoretical Probab., 7*, 73–118.

[10] Komlos, J., Major, P. and Tusnady, G. (1975) An approximation of partial sums of independent rv's and the sample df. I, *Wahrsch verw Gebiete, 32*, 111–131.

[11] Komlos, J., Major, P. and Tusnady, G. (1976) An approximation of partial sums of independent rv's and the sample df. II, *Wahrsch verw Gebiete, 34*, 33–58.

[12] Le Cam, L and Yang, G. L. (2000) *Asymptotics in Statistics*, Springer, NY.

[13] Le Cam, L. (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates, *University of California Publications in Statistics, vol 1, 11*, 277–330.

[14] Le Cam, L. (1964) Sufficiency and approximate sufficiency, *Ann. Math. Statist. 35*, 1419–1455.

[15] Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*, Springer, NY.

[16] Massart, P. (1989) Strong approximations for multivariate empirical and related processes, via KMT constructions, *Ann. Probab.* 17, 266–291.

[17] Nussbaum, M. (1996) Asymptotic equivalence of density estimation and Gaussian white noise, *Ann. Statist., 24*, 2399–2430.

[18] Steele, J. M. (2000) *Stochastic Calculus and Financial Applications.* Springer,

NY.
[19] Van der Vaart, A. W. (2002) The statistical work of Lucien Le Cam, *Ann. Statist.30*, 631–682.

ICM 2002 · Vol. I · 567–585

Ergodicity and Mixing for Stochastic Partial Differential Equations

J. Bricmont*

Abstract

Recently, a number of authors have investigated the conditions under which a stochastic perturbation acting on an infinite dimensional dynamical system, e.g. a partial differential equation, makes the system ergodic and mixing. In particular, one is interested in finding minimal and physically natural conditions on the nature of the stochastic perturbation. I shall review recent results on this question; in particular, I shall discuss the Navier-Stokes equation on a two dimensional torus with a random force which is white noise in time, and excites only a finite number of modes. The number of excited modes depends on the viscosity ν, and grows like ν^{-3} when ν goes to zero. This Markov process has a unique invariant measure and is exponentially mixing in time.

2000 Mathematics Subject Classification: 35Q30, 60H15.
Keywords and Phrases: Navier-Stokes equations with random perturbations, Markov approximations, Statistical mechanics of one-dimensional systems.

1. Introduction

The goal of this paper is to consider stochastic partial differential equations and to study conditions on the random perturbation that imply exponential convergence to a stationary state. In fact, one wants 'minimal' conditions, in the following sense: by expanding the solution in a basis of eigenfunctions of a linear operator associated with the PDE, one can write the latter as an infinite dimensional system of coupled differential equations. The question, then, is: to how many such equations do we need to add noise in order to make the system ergodic and mixing?

*UCL, Physique Théorique, Chemin du Cyclotron 2, B-1348, Louvain-la-Neuve, Belgium. E-mail: bricmont@fyma.ucl.ac.be

The physical motivation for this question comes from the fact that isotropic turbulence is often mathematically modelled by Navier Stokes equation subjected to an external stochastic driving force which is stationary in space and time. If the solution is expanded into Fourier modes, the driving force, which, in the language of physicists, acts on "large scale", should not perturb, or perturb very weakly, the high modes which represent the small scale properties of the system. So, one would like to show that the system becomes ergodic and mixing by adding noise to as few modes as possible. Obviously, this requires some detailed understanding of the nonlinear dynamics of the deterministic PDE.

This problem is interesting from another point of view. As we shall see below, one can show that all but a finite number of modes converge to equilibrium provided the remaining ones do. So, we can reduce ourselves to a finite dimensional problem, which would be standard, except for the fact that the discarded modes produce a memory effect on the remaining ones, so that the problem is no longer Markovian. At this point, one introduces techniques coming from the study of the statistical mechanics of one-dimensional systems (where the unique dimension corresponds physically to space rather than time) with "long range, exponentially decaying, interactions" which have already been very useful in the study of SRB measures in dynamical systems (see [28]).

At present, the best results require that the number of modes to which noise must be added depends on the parameters of the system, although a stronger result is likely to hold (see Remark 4 after Theorem 1.1).

The type of question discussed here (for the Navier-Stokes equation but also for other equations) has been at the center of attention of several groups of people (see Remark 3 after Theorem 1.1 below). In this paper, I shall try to explain, in a simplified form, the approach followed by A. Kupiainen, R. Lefevere and myself in [3] (see also [1, 2] for previous results).

To be concrete, consider the stochastic Navier-Stokes equation for the velocity field $u(t,x) \in \mathbf{R}^2$ defined on the torus $\mathbf{T} = (\mathbf{R}/2\pi\mathbf{Z})^2$:

$$du + ((u \cdot \nabla)u - \nu\nabla^2 u + \nabla p)dt = df \tag{1.1}$$

where $f(t,x)$ is a Wiener process with covariance

$$Ef_\alpha(t,x)f_\beta(t',y) = \min\{t,t'\}C_{\alpha\beta}(x-y) \tag{1.2}$$

and $C_{\alpha\beta}$ is a smooth function satisfying $\sum_\alpha \partial_\alpha C_{\alpha\beta} = 0$. Equation (1.1) is supplemented with the incompressibility condition $\nabla \cdot u = 0 = \nabla \cdot f$, and we will also assume that the averages over the torus vanish: $\int_{\mathbf{T}} u(0,x) = 0 = \int_{\mathbf{T}} f(t,x)$, which imply that $\int_{\mathbf{T}} u(t,x) = 0$ for all times t.

It is convenient to change (1.1) to dimensionless variables so that ν becomes equal to one. This is achieved by setting $u(t,x) = \nu u'(\nu t, x)$. Then u' satisfies (1.1), (1.2) with ν replaced by 1, and C by

$$C' = \nu^{-3}C.$$

From now on, we work with such variables and drop the primes. The dimensionless control parameter in the problem is the (rescaled) energy injection rate $\frac{1}{2}\mathrm{tr}\, C'(0)$, customarily written as $(\mathrm{Re})^3$ where Re is the Reynolds number:

$$\mathrm{Re} = \epsilon^{\frac{1}{3}}\nu^{-1},$$

and $\epsilon = \frac{1}{2}\mathrm{tr}\, C(0)$ is the energy injection rate in the original units (for explanations of the terminology see [10]).

In two dimensions, the incompressibility condition can be conveniently solved by expressing the velocity field in terms of the vorticity $\omega = \partial_1 u_2 - \partial_2 u_1$. First (1.1) implies the transport equation

$$d\omega + ((u \cdot \nabla)\omega - \nabla^2\omega)dt = db, \tag{1.3}$$

where $b = \partial_1 f_2 - \partial_2 f_1$ has the covariance

$$Eb(t,x)b(t',y) = \min\{t,t'\}(2\pi)^{-1}\gamma(x-y)$$

with $\gamma = -2\pi\nu^{-3}\Delta\mathrm{tr}C$.

Next, going to the Fourier transform, $\omega_k(t) = \frac{1}{2\pi}\int_{\mathbf{T}} e^{ik\cdot x}\omega(t,x)dx$, with $k \in \mathbf{Z}^2$; we may express u as $u_k = i\frac{(-k_2,k_1)}{k^2}\omega_k$, and write the vorticity equation as

$$d\omega(t) = F(\omega(t))dt + db(t), \tag{1.4}$$

where the drift is given by

$$F(\omega)_k = -k^2\omega_k + \frac{1}{2\pi}\sum_{l\in\mathbf{Z}^2\setminus\{0,k\}} \frac{k_1 l_2 - l_1 k_2}{|l|^2}\omega_{k-l}\omega_l \tag{1.5}$$

and $\{b_k\}$ are Brownian motions with $\bar{b}_k = b_{-k}$ and

$$Eb_k(t)b_l(t') = \min\{t,t'\}\delta_{k,-l}\,\gamma_k.$$

The dimensionless control parameter for the vorticity equation is

$$R = \sum_{k\in\mathbf{Z}^2}\gamma_k = 2\pi\gamma(0) \tag{1.6}$$

which is proportional to the ω injection rate, and also to the third power of the Reynolds number. One is interested in the turbulent region, where R is large; therefore, we will always assume below, when it is convenient, that R is sufficiently large.

For turbulence one is interested in the properties of stationary state of the stochastic equation (1.4) in the case of *smooth* forcing (see [1] for some discussion of this issue) and, ideally, one would like to consider the case where one excites only a finite number of modes,

$$\gamma_k \neq 0\ ,\ k^2 \leq N,$$

with N of order of one (for that, see Remark 4 below). In this paper we assume that N scales as

$$N = \kappa R, \tag{1.7}$$

with κ a constant, taken large enough. We set all the other $\gamma_k = 0$, although this condition can easily be relaxed. Let us denote the minimum of the covariance by

$$\rho = \min\{|\gamma_k| \mid |k|^2 \leq N\}.$$

Before stating our result, we need some definitions. Let P be the orthogonal projection in $H = L^2(\mathbf{T})$ to the subspace H_s of functions having zero Fourier components for $|k|^2 > N$. We will write

$$\omega = s + l$$

with $s = P\omega$, $l = (1-P)\omega$ (respectively, the small k and large k parts of ω). Denote also by H_l the complementary subspace (containing the nonzero components of l). H is our probability space, equipped with $\mathcal{B}$, the Borel σ-algebra.

The stochastic equation (1.4) gives rise to a Markov process $\omega(t)$ and we denote by $P^t(E|\omega)$ the transition probability of this process.

The main result of [3] is the

Theorem 1.1. *The stochastic Navier-Stokes equation* (1.4) *defines a Markov process with state space* $(H, \mathcal{B})$ *and for all* $R < \infty$, $\rho > 0$ *it has a unique invariant measure* μ *there. Moreover,* $\forall \omega \in H$, *for all Borel sets* $E \in H_s$ *and for all bounded Hölder continuous functions* F *on* H_l, *we have,*

$$|\int P^t(d\omega'|\omega) 1_E(s') F(l') - \int \mu(d\omega') 1_E(s') F(l'))| \leq C||F||_\alpha e^{-mt} \tag{1.8}$$

where $C = C(||\omega||, R, \rho) < \infty$, $m = m(R, \rho, \alpha) > 0$, *and* $||F||_\alpha$ *is the Hölder norm of exponent* α.

Remark 1. In [3], we stated, for convenience, Theorem 1.1 by saying that the constant C in (1.8) was a function of ω which was almost surely finite. Since this was stressed e.g. in [20], it is worth remarking that C is simply a function of $||\omega||$ (depending also on the parameters R and ρ), which is finite $\forall \omega \in H$. To check this, we refer the reader to equations (86) and (97) in [3]. The main reason why this bound holds, however, lies in the fact that the only dependence of our estimates on ω appears in Lemma 4.1 below and occurs through $||\omega||$.

Remark 2. In [1] it was proven that, with probability 1, the functions on the support of the measure constructed here are real analytic. In particular all correlation functions of the form

$$\int \mu(d\omega) \prod_i \nabla^{n_i} u(x_i)$$

exist. For further results on analyticity, see [26, 25].

Remark 3. While the existence of the invariant measure follows with soft methods [29], its uniqueness and the ergodic and mixing properties of the process has been harder to establish. With a nonsmooth forcing (meaning that the strength of the noise, γ_k, decays only polynomially with $|k|$) this was established in [9] and for large viscosity in [23]. However, those results did not cover the most physically interesting situations. The first result for a smooth forcing was by Kuksin and Shirikyan [13] who considered a periodically kicked system with bounded kicks (for results on exponential convergence in that model, see [14, 15, 19, 22]). In particular they could deal with the case where only a finite number of modes are excited by the noise (the number of modes depends both on the viscosity and the size of the kicks). In [2], we proved uniqueness and exponential mixing for such a kicked system where the kicks have a Gaussian distribution, but we required that there be a nonzero noise for each mode. An essential ingredient in analysis of [13], which was used in [3] and by other authors, is the Lyapunov-Schmidt type reduction that allows to transform the original Markov process with infinite dimensional state space to a non-Markovian process with finite dimensional state space. While the analysis of [13] was limited to bounded noise acting at discrete times, it was extended in [16, 17, 20, 21] to cover unbounded noise and continuous time, as well as to obtain results on the strong law of large number and the central limit theorem. The first results on ergodicity of the system with unbounded noise and finitely many excited modes were obtained in [7, 3] (see also [6] for applications to other equations) and, for exponential convergence, in [3], which was also proved in [24]. For results on related problems, see [5, 11, 12, 22].

Remark 4. What one would like to obtain is a result similar to Theorem 1.1, but with N finite, independently of R. An interesting result in that direction was obtained by Weinan E and Mattingly [8] who showed that, if one adds noise to only 2 (suitably chosen) modes, ergodicity holds, provided one truncates the system (1.4, 1.5), by keeping only a finite, but arbitrarily large, number of modes. This of course suggests that the 2 stochatically perturbed modes produce an "effective noise" on any finite number of modes, in particular on all those with $k^2 \leq \kappa R$; then, one could hope to combine this with the results in [7, 3] to obtain ergodicity and mixing for the full system. This, however, has not been done.

Remark 5. The parameters in our problem are R and ρ. All constants that do not depend on them will be generically denoted by C or c. These constants can vary from place to place.

Let me now explain the connection with ideas coming from statistical mechanics.

First, observe that, if one neglects the nonlinear term in (1.4-1.5), one expects $||\omega||$ to be of order $R^{\frac{1}{2}}$, for typical realizations of the noise ($R^{\frac{1}{2}}$ is the typical size of the noise, and the $-k^2\omega_k$ term will dominate in eq. (1.4) for larger values of $||\omega||$). It turns out that similar probabilistic estimates hold for the full equation (1.4) as shown in Section 4. Now, if $||\omega||$ is of size $R^{\frac{1}{2}}$, the $-k^2\omega_k$ term will dominate the nonlinear term (which is roughly of size $||\omega||^2$) in eq. (1.4), for $|k| \geq \kappa R^{\frac{1}{2}}$, and one

can expect that those modes (corresponding to l above) will behave somewhat like the solution of the heat equation and, in particular, that they will converge to a stationary state.

Thus, the first step is to express the l-modes in terms of the s-modes at previous times. This is done in Section 2 and produces a process for the s-modes that is no longer Markovian but has an infinite memory. In statistical mechanics, this would correspond to a system of unbounded spins (the s-modes) with infinite range interactions, with the added complications that, here, the measure is not given in a Gibbsian form, but only through a Girsanov formula, i.e. (2.9) below, and that time is continuous. Hence, we have to solve several problems: the possibility that ω be atypically large, the long range "interactions", and finally, showing that a version of the s-process with a suitable cutoff is ergodic and mixing.

In Section 3, I introduce a "toy model", namely a process with infinite memory, but with bounded variables, so that the problems caused by the unprobably large values of $||\omega||$ does not occur. In that model, I explain how the statistical mechanical techniques, developed to study systems on one dimensional lattices, can be adapted to our setting.

The large ω problem is treated in Section 4, using probabilistic estimates developed in [1], which, in statistical mechanics, would be called stability estimates. In Section 5, I sketch how the remaining problems are handled: showing that the techniques explained in Section 3 can be applied here. However, this is where several technical complications enter, for the treatment of which I refer to [3]. The problem is that, even though for typical noise, hence for typical ω's, the l-modes depend exponentially weakly on their past (see Section 2), thus producing, typically, "interactions" that decay exponentially fast, they may depend sensitively on their past when the noise is large. In the language of statistical mechanics, atypically large noise produces long range correlations, and that is the source of many technical difficulties. My goal here is to present the main conceptual tools used in [3], putting aside those difficulties.

2. Finite dimensional reduction

Using an idea of [13], one can reduce the problem of the study of a Markov process with infinite dimensional state space to that of a non-Markovian process with finite dimensional state space.

For this purpose, write the equation (1.4) for the small and large components of ω separately:

$$ds(t) = PF(s(t)+l(t))dt + db(t) \tag{2.1}$$

$$\frac{d}{dt}l(t) = (1-P)F(s(t)+l(t)). \tag{2.2}$$

The idea of [13] is to solve the l equation for a given function s, thereby defining $l(t)$ as a function of the entire history of $s(t')$, $t' \leq t$. Then, the s equation will have a drift with memory. Let us fix some notation. For a time interval I, we denote the restriction of ω (or s, l respectively) to I by $\omega(I)$, and use the boldface

notation $\mathbf{s}(I)$, to contrast it with $s(t)$, the value of s at a single time. $\|\cdot\|$ will denote the L^2 norm. In [1] it was proven that, for any $\tau < \infty$, there exists a set $\mathcal{B}_\tau$ of Brownian paths $b \in C([0,\tau], H_s)$ of full measure such that, for $b \in \mathcal{B}_\tau$, (1.4) has a unique solution with $||\omega(t)|| < \infty$, $||\nabla\omega(t)|| < \infty$ for all t (actually, $\omega(t)$ is real analytic). In particular, the projections s and l of this solution are in $C([0,\tau], H_{s(l)})$ respectively.

On the other hand, let us denote, given any $\mathbf{s} \in C([0,\tau], H_s)$, the solution — whose existence will be discussed below — of (2.2), with initial condition $l(0)$ by $l(t, \mathbf{s}([0,t]), l(0))$. More generally, given initial data $l(t')$ at time $t' < \tau$ and $\mathbf{s}([t',\tau])$, the solution of (2.2) is denoted, for $\sigma \leq \tau$, by $l(\sigma, \mathbf{s}([t',\sigma]), l(t'))$ and the corresponding ω by $\omega(\sigma, \mathbf{s}([t',\sigma]), l(t'))$. The existence and key properties of those functions are given by:

Proposition 2.1. *Let $l(0) \in H_l$ and $s \in C([0,\tau], H_s)$. Then $l(\cdot, \mathbf{s}([0,t]), l(0)) \in C([0,\tau], H_l) \cap L^2([0,\tau], H_l^1)$, where $H_l^1 = H_l \cap H^1$, and H^1 is the first Sobolev space. Moreover, given two initial conditions l_1, l_2 and $t \leq \tau$*

$$||l(t, \mathbf{s}([0,t]), l_1) - l(t, \mathbf{s}([0,t]), l_2)|| \leq \exp\left[-\kappa R t + a \int_0^t ||\nabla\omega_1||^2\right] ||l_1 - l_2|| \quad (2.3)$$

where $a = (2\pi)^{-2} \sum |k|^{-4}$ and $\omega_1(t) = s(t) + l_1(t, \mathbf{s}([0,t]), l_1)$. The solution also satisfies

$$l(t, \mathbf{s}([0,t]), l(0)) = l(t, \mathbf{s}([\tau,t]), l(\tau, \mathbf{s}([0,\tau]), l(0))). \quad (2.4)$$

Remark. What this Proposition shows is that the dependence of the function l upon its initial condition l_i, $i = 1, 2$, decays exponentially in time (i.e. like the solution of the heat equation), provided ω is not too large, in the sense that $\int_0^t ||\nabla\omega_1||^2 \leq cRt$, for a suitable constant c. As we will see in Section 4, this event is highly probable.

Now, if $s = P\omega$ with ω as above being the solution of (1.4) with noise $b \in \mathcal{B}_\tau$ then the $l(\mathbf{s})$ constructed in the Proposition equals $(1 - P)\omega$ and the stochastic process $s(t)$ satisfies the reduced equation

$$ds(t) = f(t)dt + db(t) \quad (2.5)$$

with

$$f(t) = PF(\omega(t)). \quad (2.6)$$

where $\omega(t)$ is the function on $C([0,t], H_s) \times H_l$ given by

$$\omega(t) = s(t) + l(t, \mathbf{s}([0,t]), l(0)). \quad (2.7)$$

(2.5) has almost surely bounded paths and we have a Girsanov representation for the transition probability of the ω-process in terms of the s-variables

$$P^t(F|\omega(0)) = \int \mu^t_{\omega(0)}(d\mathbf{s}) F(\omega(t)) \quad (2.8)$$

with

$$\mu^t_{\omega(0)}(d\mathbf{s}) = e^{\int_0^t (f(\tau),\gamma^{-1}(ds(\tau)-\frac{1}{2}f(\tau)d\tau))}\nu^t_{s(0)}(d\mathbf{s}) \tag{2.9}$$

where $\nu^t_{s(0)}$ is the Wiener measure with covariance γ on paths $\mathbf{s} = \mathbf{s}([0,t])$ with starting point $s(0)$ and $(\cdot,\cdot)$ the ℓ^2 scalar product. Define the operator γ^{-1} in terms of its action on the Fourier coefficients:

$$(f,\gamma^{-1}f) = \sum_{|k|^2\leq N} |f_k|^2\gamma_k^{-1}. \tag{2.10}$$

The Girsanov representation (2.8) is convenient since the problem of a stochastic PDE has been reduced to that of a stochastic process with finite dimensional state space. The drawback is that this process has infinite memory. In the next section, I will show how to deal with this problem in a simplified situation.

3. A Toy Model

In order to explain the main ideas in the proof, I will consider first a 'toy model' and then explain the steps needed to control the full model.

Let us consider variables $x_t \in [0,1]$, $t \in \mathbf{Z}$ about which a set of (consistent) conditional probability densities $p(x_t|\mathbf{x}_{[-\infty,t-1]})$ is given, i.e. one is given the probability densities of the variables x_t, at time t, given a 'past history' $\mathbf{x}_{[-\infty,t-1]}$, where we write, for $I \subset \mathbf{Z}$, $\mathbf{x}_I = (x_t)_{t\in I} \in [0,1]^I$.

Before stating precise assumptions on p, here is what one wants to prove: $\exists C < \infty$, $m > 0$ and a probability $\overline{p}$ on $[0,1]$ such that $\forall E \subset [0,1]$, E measurable,

$$|p(x_T \in E|\mathbf{x}_{[-\infty,0]}) - \overline{p}(E)| \leq Ce^{-mT} \tag{3.1}$$

for all $T > 0$ and all $\mathbf{x}_{[-\infty,0]}$, where

$$p(x_T|\mathbf{x}_{[-\infty,0]}) = \int_0^1 \prod_{t=1}^{T-1} dx_t \prod_{t=1}^{T} p(x_t|\mathbf{x}_{[1,t-1]} \vee \mathbf{x}_{[-\infty,0]}) \tag{3.2}$$

and $\mathbf{x}_{[1,t-1]} \vee \mathbf{x}_{[-\infty,0]}$ denotes the obvious configuration on $[-\infty, t-1]$.

Now let us state the assumptions on p that will imply (3.1); obviously, we assume that:

$$p(x_t|\mathbf{x}_{[-\infty,t-1]}) \geq 0 \tag{3.3}$$

and

$$\int_0^1 dx_t p(x_t|\mathbf{x}_{[-\infty,t-1]}) = 1 \tag{3.4}$$

for all $\mathbf{x}_{[-\infty,t-1]}$. Moreover, we assume that $p(\cdot|\cdot)$ is invariant under translations of the lattice $\mathbf{Z}$, in a natural way. The non-trivial assumptions are:

a) Let, for $s < t-2$,

$$\delta_{s,t}(\mathbf{x}_{[s,t]}) \equiv p(x_t|\mathbf{x}_{[s,t-1]} \vee \mathbf{0}) - p(x_t|\mathbf{x}_{[s+1,t-1]} \vee \mathbf{0}), \tag{3.5}$$

where $\mathbf{x}_I \vee \mathbf{0}$ denotes the configuration equal to x_t for $t \in I$ and equal to zero elsewhere. We assume that $\exists C < 0$, $m > 0$ such that $\forall s, t \in \mathbf{Z}$ as above,

$$\|\delta_{s,t}\|_\infty \leq C \exp(-m|t-s|). \tag{3.6}$$

b) Define, for $N \geq 1$, the Markov chain on $\Omega = [0,1]^N$ by the transition probability

$$P(\mathbf{x}_{[1,N]}|\mathbf{x}_{[-N+1,0]}) = \prod_{t=1}^{N} p(x_t|\mathbf{x}_{[t-N,t-1]} \vee \mathbf{0}). \tag{3.7}$$

We assume that this Markov chain satisfies : $\exists \delta > 0, \forall B \subset \Omega, \forall \mathbf{x}, \mathbf{x}' \in \Omega$,

$$P(B|\mathbf{x}) + P(B^c|\mathbf{x}') \geq \delta \tag{3.8}$$

where δ is independent of N (see however the Remark following the proof of Proposition 3.1 for a generalization).

Proposition 3.1. *Under assumptions a) and b) above, (3.1) holds.*

Remark 1. The techniques used here can also prove the analogue of (3.1) with x_T replaced by $x_{[T,T-L]}$, for any finite L, and this, in turn, allows one to associate to the system of conditional probabilities a unique probability distribution on $[0,1]^{\mathbf{Z}}$ (which is called, in statistical mechanics, the Gibbs state associated to the system of conditional probabilities), but I will not go into that, because I want to give here only an elementary idea of the techniques used in [3]. Of course, this type of results is not new (see e.g. [28], Lecture 12, for a similar result, applied to dynamical systems, with a somewhat different proof).

To prove the Proposition, we first use a result of Doob ([4], p. 197–198):

Lemma 3.1 *For the Markov chain defined in b) above, there exists a probability distribution P on Ω such that $\forall \mathbf{x} \in \Omega$, $\forall B \subset \Omega$, $\forall n \geq 1$,*

$$|P^n(B|\mathbf{x}) - P(B)| \leq (1-\delta)^n. \tag{3.9}$$

Proof. Let $\overline{P}(B,n) = \sup_{\mathbf{x}} P^n(B|\mathbf{x})$ and $\underline{P}(B,n) = \inf_{\mathbf{x}} P^n(B|\mathbf{x})$. It is easy to see that $\overline{P}(B,n)$ is decreasing in n, while $\underline{P}(B,n)$ is increasing in n. Thus, it is sufficient to prove the bound (3.9) for the difference $|\overline{P}(B,n) - \underline{P}(B,n)|$ and, for that, we shall prove:

$$0 \leq \overline{P}(B,n+1) - \underline{P}(B,n+1) \leq (1-\delta)(\overline{P}(B,n) - \underline{P}(B,n)). \tag{3.10}$$

Since $\overline{P}(B,n) - \underline{P}(B,n) \leq 1$, (3.9) follows.

Define a signed measure on subsets of Ω:

$$\Psi_{\mathbf{x},\mathbf{x}'}(E) = P(E|\mathbf{x}) - P(E|\mathbf{x}') \tag{3.11}$$

and let S^+ (resp. S^-) denote the set where $\Psi_{\mathbf{x},\mathbf{x}'}(E) \geq 0$ for $E \subset S^+$ (resp. ≤ 0).

We have:

$$\overline{P}(B, n+1) - \underline{P}(B, n+1) = \sup_{\mathbf{x},\mathbf{x}'} \int [P(d\mathbf{x}''|\mathbf{x}) - P(d\mathbf{x}''|\mathbf{x}')] P^n(B|\mathbf{x}'')$$

$$= \sup_{\mathbf{x},\mathbf{x}'} \int \Psi_{\mathbf{x},\mathbf{x}'}(d\mathbf{x}'') P^n(B|\mathbf{x}'')$$

$$\leq \sup_{\mathbf{x},\mathbf{x}'} (\Psi_{\mathbf{x},\mathbf{x}'}(S^+)\overline{P}(B,n) + \Psi_{\mathbf{x},\mathbf{x}'}(S^-)\underline{P}(B,n)). \tag{3.12}$$

By definition, $\Psi_{\mathbf{x},\mathbf{x}'}(S^-) = -\Psi_{\mathbf{x},\mathbf{x}'}(S^+)$, so that

$$\Psi_{\mathbf{x},\mathbf{x}'}(S^+)\overline{P}(B,n) + \Psi_{\mathbf{x},\mathbf{x}'}(S^-)\underline{P}(B,n) = \Psi_{\mathbf{x},\mathbf{x}'}(S^+)(\overline{P}(B,n) - \underline{P}(B,n)).$$

Also, for any set $E \subset \Omega$, (3.8) implies

$$\Psi_{\mathbf{x},\mathbf{x}'}(E) = 1 - (P(E^c|\mathbf{x}) + P(E|\mathbf{x}')) \leq 1 - \delta.$$

Applying this to $E = S^+$ in (3.12) implies (3.10).

Remark 2. We shall use this Lemma under the following form:

$$\int_\Omega d\mathbf{x} |P^n(\mathbf{x}|\mathbf{x}') - P(\mathbf{x})| \leq 2(1-\delta)^n \tag{3.13}$$

for all $\mathbf{x}' \in \Omega$; this follows by applying (3.9) separately to the sets where the integrand is positive and negative.

Now, let us turn to the

Proof of Proposition 3.1.

We write each factor in (3.2) as

$$p(x_t|\mathbf{x}_{[-\infty,t-1]}) = p(x_t|\mathbf{x}_{[t-N,t-1]} \vee \mathbf{0}) + \sum_{|s-t|>N} \delta_{s,t}(\mathbf{x}_{[s,t]}), \tag{3.14}$$

where N is an integer to be chosen later. Insert this in the product in (3.2), and expand: we get

$$p(x_T|\mathbf{x}_{[-\infty,0]}) = \sum_{I \subset [1,T]} \sum_{\mathbf{s}} \int \prod_{t=1}^{T-1} dx_t \prod_{t \in I} \delta_{s,t}(\mathbf{x}_{[s,t]}) \prod_{t \notin I} p(x_t|\mathbf{x}_{[t-N,t-1]} \vee \mathbf{0}), \tag{3.15}$$

where the sum over subsets I corresponds to the choice in (3.14) between the first term and the sum, while the sum over $\mathbf{s} = (s_t)_{t \in I}$ corresponds to the possible choices of a term in that sum.

Now, let $\overline{I} = \bigcup_{t\in I}[s_t, t]$ and let

$$[1,T]\backslash\overline{I} = \bigcup_i J_i \bigcup_\alpha I_\alpha, \tag{3.16}$$

where each J_i is a union of intervals of length N, containing at least two such intervals, and each I_α is an interval of length less than $2N$ between two connected intervals in $\overline{I}$ or an interval of length less than N between an interval in $\overline{I}$ and an interval J_i. The reason for these definitions is that, in the RHS of (3.15), the only functions depending on x_s, with s in the complement of $\overline{I}$, are the factors $p(x_t|\mathbf{x}_{[t-N,t-1]})$, so that, by integrating over these variables, one can obtain the transition probabilities of the Markov chain defined in condition b) above. For that, we need intervals of length at least $2N$, which are the J_i's, while the intervals I_α's simply cover the leftover sites.

Since the model here is translation invariant, let us fix one interval $J_i = J$, and write it as a union of disjoint intervals of length N: $J = \bigcup_{l=0}^{n} K_l$ with $K_l = [t+1+lN, t+(l+1)N]$.

We have, by definition (3.7) of the transition probability P:

$$\int \prod_{l=1}^{n-1} d\mathbf{x}_{K_l} \prod_{l=1}^{n} \prod_{t\in K_l} (p(x_t|\mathbf{x}_{[t-N,t-1]} \vee \mathbf{0})) = P^n(\mathbf{x}_{K_n}|\mathbf{x}_{K_0}). \tag{3.17}$$

Now write this as

$$P^n(\mathbf{x}_{K_n}|\mathbf{x}_{K_0}) - P(\mathbf{x}_{K_n}) + P(\mathbf{x}_{K_n}), \tag{3.18}$$

where P is defined by (3.9). Apply this to each interval J_i in (3.16), with n replaced by $n_i = \frac{|J_i|}{N} - 1$. Insert that identity in (3.15) for each J_i and expand the corresponding product over i of $A_i + B_i$, where $A = P^n(\mathbf{x}_{K_n}|\mathbf{x}_{K_0}) - P(\mathbf{x}_{K_n})$ and $B = P(\mathbf{x}_{K_n})$.

For E as in (3.1), integrate over E each term in the resulting expansion, and write

$$p(x_T \in E|\mathbf{x}_{[-\infty,0]}) = Q + R, \tag{3.19}$$

where Q collects all the terms in the resulting sum where at least one factor $P(\mathbf{x}_{K_{n_i}})$ appears and R all the rest. Now, the presence of one such factor P 'decouples' x_T from the initial conditions $\mathbf{x}_{[-\infty,0]}$, in the sense that, if we consider the difference

$$p(x_T \in E|\mathbf{x}_{[-\infty,0]}) - p(x_T \in E|\mathbf{x}'_{[-\infty,0]}), \tag{3.20}$$

for two different past histories, then the Q sums are equal and only the R sums contribute to the difference. Indeed, fix a K_{n_i} and consider all the terms in our expansions where the factor $P(\mathbf{x}_{K_{n_i}})$ appears; let t_0 be the last time before the interval K_{n_i}. By construction, in all the terms under consideration, all the functions

that depend on x_t, for $t > t_0$ do not depend on the variables x_t, for $t \leq t_0$. So, if we resum, in the expansion, all the terms depending on the variables x_t, for $t \leq t_0$, we obtain, for the two terms in (3.20), $p(x_{t_0}|\mathbf{x}_{[-\infty,0]})$, and $p(x_{t_0}|\mathbf{x}'_{[-\infty,0]})$ (we simply use (3.15) read from right to left, with T replaced by t_0). But performing in (3.2) the integral over x_{t_0} gives 1 in both cases, which shows that the difference between the respective sums cancel.

So, if we show that, $\exists C < \infty, m > 0$ such that

$$|R| \leq Ce^{-mT}, \tag{3.21}$$

$\forall \mathbf{x}_{[-\infty,0]}$, we obtain that the absolute value of (3.20) is exponentially small and, from that, (3.1) easily follows.

Using the bound (3.6) on $\delta_{s,t}$ and (3.13) on

$$\int_\Omega \prod_{t \in K_n} dx_t |P^n(\mathbf{x}_{K_n}|\mathbf{x}_{K_0}) - P(\mathbf{x}_{K_n})|, \tag{3.22}$$

and the fact that, by (3.4) and $x_t \in [0,1]$, all the integrals are bounded by 1, , we get:

$$|R| \leq \sum_I \sum_{\mathbf{s}} \prod_{t \in I} (Ce^{-m|t-s_t|}) \prod_i (2(1-\delta)^{n_i}), \tag{3.23}$$

where the second product runs over the intervals J_i in (3.16), and where $n_i = \frac{|J_i|}{N} - 1$. Note that the length of each I_α in (3.16) is less than $2N$ and, since such intervals are always adjacent to a connected component of $\overline{I}$ (unless $I = \emptyset$, in which case this number is at most 2), the number of intervals I_α is less than $2|I| + 2$; the same bound holds for the number of intervals J_i in (3.16) (in fact, a better bound holds here, but we won't use it). So, we have:

$$\sum_i n_i \geq \sum_i \frac{|J_i|}{N} - (2|I| + 2) \geq \frac{(T - |\overline{I}|)}{N} - c|I| - 2, \tag{3.24}$$

for some number c, where, in the second inequality, we use $|I_\alpha| \leq 2N$ and (3.16).

Using this, we can, by changing the constant C, bound (3.23) by:

$$C \sum_I \sum_{\mathbf{s}} \prod_{t \in I} e^{-m|t-s_t|} C^{|I|} (1-\delta)^{(T-|\overline{I}|)/N} \tag{3.25}$$

Since by definition of $\overline{I}$, $\sum_{t \in I} |t - s_t| \geq |\overline{I}|$, we can, by considering separately the terms where $|\overline{I}| \leq \frac{T}{2}$, and those where $|\overline{I}| > \frac{T}{2}$, bound the sum in (3.25) by

$$Ce^{-\tilde{m}T} \sum_I \sum_{\mathbf{s}} \prod_{t \in I} e^{-\frac{m}{2}|t-s_t|} C^{|I|} \tag{3.26}$$

where

$$\tilde{m} = \min\left(\frac{m}{4}, \frac{-\ln(1-\delta)}{2N}\right). \tag{3.27}$$

Now, choose N so that

$$\sum_{|t-s|>N} e^{-\frac{m}{2}|t-s|} \leq \eta, \tag{3.28}$$

with

$$(1 + C\eta) \leq e^{\tilde{m}/2}, \tag{3.29}$$

which is possible since, from (3.28) we see that, for large N, $\eta = \exp(-\mathcal{O}(N))$ while, from (3.27), $\tilde{m} = \mathcal{O}(N^{-1})$.

We use (3.28) to control the sum over each s_t in (3.26), and we get

$$(3.26) \leq Ce^{-\tilde{m}T} \sum_{I \subset [1,T]} (C\eta)^{|I|} \leq C(1 + C\eta)^T e^{-\tilde{m}T} \tag{3.30}$$

and, using (3.29), we get (3.21) with $m = \frac{\tilde{m}}{2}$.

This completes the proof of Proposition 3.1.

Remark 3. By considering (3.27, 3.28, 3.29), we see that one can extend the proof to a situation where δ in (3.8) depends on N, as long as $\delta \geq \exp(-cN)$ for a constant c small enough.

Now, let us turn to the real model, and make a list of the difficulties not present in the toy model. The first one is that time is continuous rather than discrete, but that is a minor problem. We can easily introduce a discretization of time. A more serious problem is that one deals with what are called "unbounded spins" in statistical mechanics or what is also known as a "large field problem", namely the variables $s(\tau)$ in (2.9), which play a role similar to the variables x_t here, take value in $\mathbf{R}^N$ rather than $[0, 1]$ (actually, if we consider the variable s over a unit time interval, they take values in a space of functions from that interval into $\mathbf{R}^N$). And, what really causes a problem, is the fact that the bounds (3.6), (3.8) do not hold when the variables s take large values. However, as we shall see in the next section, this is unprobable. Thus, before doing an expansion as in (3.15, 3.18), we must first distinguish between time intervals where the s variables are large and those where they are small. Then, putting aside lots of technicalities, we perform the expansion (3.15) in the latter intervals and use estimates like (4.2) below to control the sum over the intervals where ω is large.

Finally, there is an additional difficulty coming from the fact that the definition of the probabilities here involve a Girsanov representation. In statistical mechanics, one usually deals with situations where the probabilities (3.3) can be written as:

$$p(x_t|\mathbf{x}_{[-\infty,t-1]}) = \exp(\sum_{t \in I} \phi_I(x_I)), \tag{3.31}$$

where the $\{\phi_I\}$'s represent "many body interactions' (suitably normalized so that (3.4) holds) and the sum runs over intervals $I \subset \mathbf{Z}$ whose last point is t. Then, a bound of the form

$$\|\phi_I\|_\infty \leq C \exp(-m|I|), \tag{3.32}$$

with $C < \infty$, $m > 0$, is enough to obtain (3.6) and (3.8). But here the probabilities are not of that form, because of the stochastic integral $\int_0^t f(\tau)\gamma^{-1}ds(\tau)$ in (2.9).

4. A priori estimates on the transition probabilities

The memory in the process (2.5) is coming from the dependence of the solution of (2.2) on its initial conditions. By Proposition 2.1, the dependence is weak if $\int_0^t ||\nabla\omega||^2$ is less than cRt for a suitable c. It is convenient to define, for each unit interval $[n-1, n] \equiv \mathbf{n}$, a quantity measuring the size of ω on that interval by:

$$D_n = \frac{1}{2}\sup_{t\in\mathbf{n}} ||\omega(t)||^2 + \int_{\mathbf{n}} ||\nabla\omega(t)||^2 dt. \tag{4.1}$$

The following Proposition bounds the probability of the unlikely event that we are interested in:

Proposition 4.1. *There exist constants $c > 0$, $c' < \infty$, $\beta_0 < \infty$, such that for all t, t', $1 \le t < t'$ and all $\beta \ge \beta_0$,*

$$P\Big(\sum_{n=t}^{t'-1} D_n(\omega) \ge \beta R|t'-t|\Big|\omega(0)\Big) \le \exp(\tfrac{1}{R}c'e^{-t}||\omega(0)||^2)\exp(-c\beta|t'-t|). \tag{4.2}$$

Remark 1. This means that the probability that ω is large over an interval of time decays exponentially with the length of that interval, provided that $||\omega(0)||$ is not too large. And, if $||\omega(0)||^2$ is of order K, $D_n(\omega)$ will be, with large probability, of order R after a time of order $\log K$.

The main idea in the proof is a probabilistic analogue of the so-called enstrophy balance: in the deterministic case, using integration by parts and $\nabla \cdot u = 0$, on derives from (1.3) with $db = 0$, the identity:

$$\frac{1}{2}\frac{d}{dt}||\omega||^2 = -||\nabla\omega||^2,$$

which implies that the enstrophy ($||\omega||^2$) decreases in time. This basic property of equation (1.3) makes the proof of the following Lemma rather simple.

Lemma 4.1. *For all $\omega(0) \in L^2$, and all $t \ge 0$,*

$$E\Big[e^{\frac{1}{4R}||\omega(t)||^2}\ \Big|\ \omega(0)\Big] \le 3e^{\frac{1}{4R}e^{-t}||\omega(0)||^2}, \tag{4.3}$$

and

$$P(||\omega(t)||^2 \ge D|\omega(0)) \le 3e^{-\frac{D}{4R}}e^{\frac{1}{4R}e^{-t}||\omega(0)||^2}. \tag{4.4}$$

Proof. Let $x(\tau) = \lambda(\tau)\|\omega(\tau)\|^2 = \lambda(\tau)\sum_k |\omega_k|^2$ for $0 \le \tau \le t$. Then by Ito's formula (remember that, by (1.6), $\sum_k \gamma_k = R$ and thus $\gamma_k \le R$, $\forall k$):

$$\begin{aligned} \frac{d}{d\tau}E[e^x] &= E[(\dot\lambda\lambda^{-1}x - 2\lambda\sum_k k^2|\omega_k|^2 + \lambda\sum_k \gamma_k + 2\lambda^2\sum_k \gamma_k|\omega_k|^2)e^x] \\ &\le E[((\dot\lambda\lambda^{-1} - 2 + 2\lambda R)x + \lambda R)e^x], \end{aligned} \tag{4.5}$$

where E denotes the conditional expectation, given $\omega(0)$, and where we used the Navier-Stokes equation (1.3), $|k| \ge 1$ for $\omega_k \ne 0$, and the fact that the nonlinear term does not contribute (using integration by parts and $\nabla \cdot u = 0$). Take now $\lambda(\tau) = \frac{1}{4R}e^{(\tau - t)}$ so that $\lambda \le \frac{1}{4R}$, $\dot\lambda\lambda^{-1} = 1$, $\dot\lambda\lambda^{-1} - 2 + 2\lambda R \le -\frac{1}{2}$ and $\lambda R \le \frac{1}{4}$. So,

$$\frac{d}{d\tau}E[e^x] \le E[(\frac{1}{4} - \frac{1}{2}x)e^x] \le \frac{1}{2} - \frac{1}{4}E[e^x],$$

where the last inequality follows by using $(1 - 2x)e^x \le 2 - e^x$. Thus, Gronwall's inequality implies that:

$$E[e^{x(\tau)}] \le e^{-\frac{\tau}{4}}e^{x(0)} + 2 \le 3e^{x(0)},$$

i.e., using the definition of $\lambda(\tau)$,

$$E\Big[\exp(\frac{e^{\tau - t}}{4R}\|\omega(\tau)\|^2)\Big] \le 3\exp(\frac{e^{-t}\|\omega(0)\|^2}{4R}).$$

This proves (4.3) by putting $\tau = t$; (4.4) follows from (4.3) by Chebychev's inequality.

Since the D_n in (4.2) is the supremum over unit time intervals of

$$D_t(\omega) = \frac{1}{2}\|\omega(t)\|^2 + \int_{n-1}^{t} \|\nabla\omega\|^2 d\tau \qquad n - 1 \le t \le n, \tag{4.6}$$

which does not involve only $\|\omega(t)\|^2$, we need to control also the evolution of $D_t(\omega)$ over a unit time interval, taken, for now, to be $[0, 1]$. From the Navier-Stokes equation (1.3) and Ito's formula, we obtain

$$D_t(\omega) = D_0(\omega) + Rt + \int_0^t (\omega, db) \tag{4.7}$$

(since the nonlinear term does not contribute, as in (4.5)).

Our basic estimate is:

Lemma 4.2. *There exist $C < \infty$, $c > 0$ such that, $\forall A \ge 3D_0(\omega)$*

$$P(\sup_{t\in[0,1]} D_t(\omega) \ge A|\omega(0)) \le C\exp(-\frac{cA}{R}). \tag{4.8}$$

Remark 2. While the previous Lemma showed that $||\omega(t)||^2$ tends to decrease as long as it is larger than $\mathcal{O}(R)$, this Lemma shows that, in a unit interval, $D_t(\omega)$ does not increase too much relative to $D_0(\omega) = \frac{1}{2}||\omega(0)||^2$. Thus, by combining these two Lemmas, we see that $D_n(\omega) = \sup_{t\in[n-1,n]} D_t(\omega)$ is, with large probability, less than $||\omega(0)||^2$, when the latter is larger than $\mathcal{O}(R)$, at least for $n \geq n_0$ not too small. Thus, it is unlikely that $D_n(\omega)$ remains much larger than R over some interval of (integer) times, and this is the basis of the proof of Proposition 4.1.

Without entering into details, here are the main ideas in the proof of (4.8). From (4.7), we see that it is enough to get an upper bound on

$$P\left(\sup_{t\in[0,1]} |\int_0^t (\omega, db)| \geq (A - D_0 - R)\Big|\omega(0)\right). \tag{4.9}$$

We use (see [3] for more details) Doob's inequality (see e.g.[27], p.24), to reduce the control over the supremum over t to estimates on $|\int_0^1 (\omega, db)|$. Letting E denote the conditional expectation, given $\omega(0)$. and using Novikov's bound (see e.g. the proof of Lemma 5.2 below), we get

$$\begin{aligned} E(e^{\pm\varepsilon\int_0^1(\omega,db)}) &\leq \left(E(e^{2\varepsilon^2\int_0^1 d\tau(\omega(\tau),\gamma\omega(\tau))})\right)^{1/2} \\ \leq \left(\int_0^1 d\tau E(e^{2\varepsilon^2(\omega(\tau),\gamma\omega(\tau))})\right)^{1/2} &\leq \left(\int_0^1 d\tau E(e^{2\varepsilon^2 R||\omega(\tau)||^2})\right)^{1/2} \end{aligned} \tag{4.10}$$

where the last two inequalities follow from Jensen's inequality, applied to $e^{2\varepsilon^2\int_0^1 d\tau(\omega(\tau),\gamma\omega(\tau))}$, and from $\gamma_k \leq R$ (see (1.6)). Now, choosing ε so that $2\varepsilon^2 R = \frac{1}{4R}$, i.e. $\varepsilon = \frac{1}{\sqrt{8R}}$, we can use (4.4) to bound the RHS of (4.10). Combining this with Chebychev's inequality gives bounds on (4.9).

5. Decoupling estimates

In this section, I shall give a very brief sketch of the ideas used to prove the analogue of assumptions (3.6) and (3.8) of section 3 in the present setting, at least in the probable regions where ω is small. The main point is to understand the analogue of the bound (3.32), which expresses the exponential decay of interactions. What plays the role of the right hand side of (3.31) is, see (2.9):

$$g_t \equiv e^{\int_{t-1}^t (f(\tau),\gamma^{-1}(ds(\tau)-\frac{1}{2}f(\tau)d\tau))} \tag{5.1}$$

where, for simplicity, I consider a unit time interval $[t-1, t]$. We want to show that this depends weakly on the past; so consider two functions g_1, g_2, defined in terms of two functions f_1, f_2, themselves defined through different l_1 and l_2 (see (2.6, 2.7)). And, by analogy with what we did in section 3, we choose $l_1 = l(t, \mathbf{s}([0,t]), l(0) = 0)$, $l_2 = l(t, \mathbf{s}([1,t]), l(1) = 0)$, i.e. we set the large k modes equal to zero at different times (0 or 1). Using (2.4), we see that $l_1 = l(t, \mathbf{s}([1,t]), l_1(1))$,

with $l_1(1) = l_1(1, \mathbf{s}([0,1]), l(0) = 0)$, so that we have, at time $t = 1$, two initial conditions, $l_1(1)$, $l_2(1) = 0$, with $\|l_1(1) - l_2(1)\| = \|l_1(1)\|$ of order one, if ω is small in the interval $[0,1]$.

Now, if g_t depends weakly on the past, it should mean that, for large t, g_1 and g_2 are, in some sense, exponentially close. To measure the difference, write:

$$g_1 - g_2 = (1 - \frac{g_2}{g_1})g_1 \equiv (1 - H)g_1, \tag{5.2}$$

which will be convenient, since we deal with unbounded variables for which sup norm estimates like in (3.32) are not available. Explicitly:

$$H = e^{\int_{t-1}^{t}(\delta f(t), \gamma^{-1}(ds(t) - f_1(t)dt)) - \frac{1}{2}\int_{t-1}^{t}(\delta f(t), \gamma^{-1}\delta f(t))dt} \tag{5.3}$$

where $\delta f = f_2 - f_1$. What we want to show is that $1 - H$ is, in a suitable sense, small.

The next Lemma gives a bound on $\|\delta f\|$ in terms of $\|\delta l\|$, and $\|\omega\|$; $\|\delta l\|$ is controlled by Proposition 2.1, provided that ω is small, in the sense discussed in section 4, in which case $\|\omega\|$ is also controlled, using $\sup_{t \in \mathbf{n}} \|\omega(t)\| \leq (2D_n)^{\frac{1}{2}}$.

Lemma 5.1. *Let $f(\omega) = PF(\omega)$ and $\omega = s + l$, $\omega' = s + l'$. Then,*

$$\|\delta f\| = \|f(\omega) - f(\omega')\| \leq C(R)(2\|\omega\|\|\delta l\| + \|\delta l\|^2) \tag{5.4}$$

with $\delta l = l - l'$ and $C(R)$ a constant depending on the parameter R (see (1.6)).

Proof. We have

$$|f_k(\omega) - f_k(\omega')| \leq \sum_p |\omega_{k-p}\omega_p - \omega'_{\kappa-p}\omega'_p| \frac{|k|}{|p|}$$

which, since $|k| \leq \sqrt{\kappa R}$ is bounded by

$$\sqrt{\kappa R} \sum_p |s_{k-p}\delta l_p + s_p \delta l_{k-p} + l_p l_{k-p} - l'_p l'_{k-p}|. \tag{5.5}$$

Writing $l_p l_{k-p} - l'_p l'_{k-p} = l_p \delta l_{k-p} + l_{k-p}\delta l_p - \delta l_p \delta l_{k-p}$ and using Schwarz' inequality, we get

$$(5.5) \leq \sqrt{\kappa R}(2\|\omega\|\|\delta l\|\| + \|\delta l\|^2)$$

which proves (5.4), since $f_k \neq 0$ only for $k \leq \kappa R$, so that the sum in the L^2 norm $\|\delta f\|$ runs over $C(R)$ terms.

This Lemma would be enough to control $(1 - H)$ if we had only in (5.3) the factor $e^{-\frac{1}{2}\int_{t-1}^{t}(\delta f(t), \gamma^{-1}\delta f(t))dt}$, which involves only ordinary integrals.

To control the stochastic integral, it is convenient to undo the Girsanov transformation, i.e. to change variables from s back to b. Let E denote the expectation

with respect to the Brownian motion b with covariance γ on the time interval $[t-1,t]$. We get, using (2.5):

$$H = e^{\int_{t-1}^{t}(\delta f(t),\gamma^{-1}db(t))-\frac{1}{2}\int_{t-1}^{t}(\delta f(t),\gamma^{-1}\delta f(t))dt}. \tag{5.6}$$

Write now $(1-H)^2 = 1-2H+H^2$; to give a flavour of the estimates, let us see how one could show that the expectation with respect to E of $-2H+H^2$is close to 1, i.e. that the expectation of $(1-H)^2$ is close to zero. One can rather easily bound from below the expectation of H, using Jensen's inequality; to get an upper bound on the expectation of H^2, one uses:

Lemma 5.2. *Let $\zeta(t) \in C([0,1],H_s)$ be progressively measurable. Then*

$$Ee^{\int_0^1(\zeta,\gamma^{-1}db)+\lambda\int_0^1(\zeta,\gamma^{-1}\zeta)dt} \leq e^{2(1+\lambda)||\zeta||^2\rho^{-1}} \tag{5.7}$$

where $||\zeta|| = \sup_\tau ||\zeta(\tau)||_2$.

Proof. This is just a Novikov bound: we bound the LHS, using Schwarz' inequality, by

$$(Ee^{\int_0^1(2\zeta,\gamma^{-1}db)-2\int_0^1(\zeta,\gamma^{-1}\zeta dt)})^{\frac{1}{2}}(Ee^{2(1+\lambda)\int_0^1(\zeta,\gamma^{-1}\zeta)dt})^{\frac{1}{2}}$$

and note that the expression inside the first square root is the expectation of a martingale and equals one.

We can then apply this Lemma to $\zeta = 2\delta f$, $\lambda = -\frac{1}{4}$, replacing $[0,1]$ by $[t-1,t]$, and use the estimates coming from Lemma 5.1 and Proposition 2.1 to show that the RHS of (5.7) is exponentially close to 1, for t large. This gives a rough idea of why the "interactions" here are exponentially decaying, but it must be said that the full story is far more complicated and I refer to reader to [3] for more details.

Acknowledgments.

I wish to thank my coauthors, Antti Kupiainen and Raphael Lefevere without whom this work would not have been possible. This work was supported in part by ESF/PRODYN.

References

[1] J. Bricmont, A. Kupiainen, R. Lefevere, Probabilistic estimates for the two dimensional stochastic Navier-Stokes equations. *J. Stat. Phys.* **100** (3/4), (2000), 743–756.

[2] J. Bricmont, A. Kupiainen, R. Lefevere, Ergodicity of the 2D Navier-Stokes Equations with random forcing. *Commun. Math. Phys.* **224** (2001), 65–81.

[3] J. Bricmont, A. Kupiainen, R. Lefevere, Exponential mixing of the 2D stochastic Navier-Stokes dynamics *Commun. Math. Phys.* (to appear).

[4] J.L. Doob, *Stochastic Processes*, John Wiley, 1953.

[5] J.P. Eckmann, M. Hairer, Uniqueness of the invariant measure for a stochastic PDE driven by degenerate noise, *Commun. Math. Phys.* **219** (2001), 523–565.

[6] W. E, D. Liu, Gibbsian dynamics and invariant measures for stochastic dissipative PDEs, *J. Stat. Phys.* (to appear).

[7] W. E, J.C. Mattingly, Ya.G. Sinai, Gibbsian dynamics and ergodicity for the stochastically forced Navier-Stokes equation, *Comm. Math. Phys.* **224** (2001), 83–106.
[8] W. E, J.C. Mattingly, Ergodicity for the Navier-Stokes equation with degenerate random forcing, finite-dimensional approximation, *Comm. on Pure and Appl. Math.* **LIV** (2001), 1386–1402.
[9] F. Flandoli, B. Maslowski, Ergodicity of the 2-D Navier-Stokes equation under random perturbations. *Commun. Math. Phys.* **171** (1995), 119–141.
[10] U. Frisch, *Turbulence*, Cambridge University Press, 1995.
[11] M. Hairer, Exponential mixing properties of stochastic PDE's through asymptotic coupling, Preprint (2001).
[12] M. Hairer, Exponential mixing for a stochastic PDE driven by degenerate noise *Nonlinearity* **15** (2002), 271–279.
[13] S.B. Kuksin, A. Shirikyan, Stochastic dissipative PDE's and Gibbs measures. *Commun. Math. Phys.* **213** (2000), 291–330.
[14] S.B. Kuksin, A. Shirikyan, A coupling approach to randomly forced nonlinear PDE's. I, *Comm. Math. Phys.* **221** (2001), 351–366.
[15] S.B. Kuksin, A. Piatnitski, A. Shirikyan, A coupling approach to randomly forced nonlinear PDE's. II, *Comm. Math. Phys.* (to appear).
[16] S.B. Kuksin, A. Shirikyan, Ergodicity for the randomly forced 2D Navier-Stokes equations, *Math. Phys. Anal. Geom.* **4** (2001), 147–195.
[17] S.B. Kuksin, A. Shirikyan, Coupling approach to white-forced nonlinear PDE's, *Journ. de Math. Pures et Appl.*, **81** (2002), 567–602.
[18] S.B. Kuksin, A. Shirikyan, On dissipative systems perturbed by bounded random kick-forces *Ergodic Theory and Dynamical Systems* (to appear).
[19] S.B. Kuksin, On exponential convergence to a stationary measure of nonlinear PDEs, perturbed by random kick-forces, and the turbulence-limit, *The M.I. Vishik Moscow PDE seminar, Amer. Math. Soc. Transl.*, (to appear).
[20] S.B. Kuksin, Ergodic theorems for 2D statistical hydrodynamics, *Rev. Math. Phys.* (to appear).
[21] S.B. Kuksin, A. Shirikyan, Some limiting properties of randomly forced 2D Navier-Stokes equations, preprint (2002).
[22] N. Masmoudi, L.-S. Young, Ergodic theory of infinite dimensional systems with applications to dissipative parabolic PDE's. Preprint (2001).
[23] J.C. Mattingly, Ergodicity of 2D Navier-Stokes equations with random forcing and large viscosity, *Commun. Math. Phys.* **206** (1999), 273–288.
[24] J. C. Mattingly, Exponential convergence for the stochastically forced Navier-Stokes equations and other partially dissipative dynamics, preprint (2001).
[25] J. C. Mattingly, The dissipative scale of the stochastic Navier-Stokes equation: regularization and analyticity, *J. Stat. Phys* (to appear).
[26] A. Shirikyan, Analyticity of solutions for randomly forced two-dimensional Navier-Stokes equations, *Russian Math. Surveys*, (to appear).
[27] B. Simon, *Functional Integration and Quantum Physics*, Academic Press, 1979.
[28] Ya.G. Sinai, *Topics in Ergodic Theory*, Princeton University Press, 1994.
[29] M.I., Vishik, A.V. Fursikov, *Mathematical Problems of Statistical Hydrodynamics*, Kluwer, 1980.

ICM 2002 · Vol. I · 587–596

Distribution Functions for Largest Eigenvalues and Their Applications

Craig A. Tracy* Harold Widom†

Abstract

It is now believed that the limiting distribution function of the largest eigenvalue in the three classic random matrix models GOE, GUE and GSE describe new universal limit laws for a wide variety of processes arising in mathematical physics and interacting particle systems. These distribution functions, expressed in terms of a certain Painlevé II function, are described and their occurences surveyed.

2000 Mathematics Subject Classification: 82D30, 60K37.
Keywords and Phrases: Random matrix models, Limit laws, Growth processes, Painlevé functions.

1. Random matrix models

A random matrix model is a probability space $(\Omega, \mathcal{P}, \mathcal{F})$ where the sample space Ω is a set of matrices. There are three classic finite N random matrix models (see, e.g. [31] and for early history [37]):

- Gaussian Orthogonal Ensemble ($\beta = 1$)
  - $\Omega = N \times N$ real symmetric matrices
  - $\mathcal{P}$ = "unique" measure that is invariant under orthogonal transformations and the matrix elements are i.i.d. random variables. Explicitly, the density is
$$c_N \exp\left(-\mathrm{tr}(A^2)\right) dA, \tag{1.1}$$
where c_N is a normalization constant and $dA = \prod_i dA_{ii} \prod_{i<j} dA_{ij}$, the product Lebesgue measure on the independent matrix elements.
- Gaussian Unitary Ensemble ($\beta = 2$)
  - $\Omega = N \times N$ hermitian matrices

*Department of Mathematics, University of California, Davis, CA 95616, USA. E-mail: tracy@math.ucdavis.edu
†Department of Mathematics, University of California, Santa Cruz, CA 95064, USA. E-mail: widom@math.ucsc.edu

  - $\mathcal{P}=$ "unique" measure that is invariant under unitary transformations and the (independent) real and imaginary matrix elements are i.i.d. random variables.
- Gaussian Symplectic Ensemble ($\beta = 4$) (see [31] for a definition)

Generally speaking, the interest lies in the $N \to \infty$ limit of these models. Here we concentrate on one aspect of this limit. In all three models the eigenvalues, which are random variables, are real and with probability one they are distinct. If $\lambda_{\max}(A)$ denotes the largest eigenvalue of the random matrix A, then for each of the three Gaussian ensembles we introduce the corresponding distribution function

$$F_{N,\beta}(t) := \mathrm{P}_\beta\left(\lambda_{\max} < t\right), \beta = 1, 2, 4.$$

The basic limit laws [46, 47, 48] state that[1]

$$F_\beta(s) := \lim_{N\to\infty} F_{N,\beta}\left(2\sigma\sqrt{N} + \frac{\sigma s}{N^{1/6}}\right), \ \beta = 1, 2, 4,$$

exist and are given explicitly by

$$\begin{aligned} F_2(s) &= \det\left(I - K_{\mathrm{Airy}}\right) \\ &= \exp\left(-\int_s^\infty (x-s)q^2(x)\,dx\right) \end{aligned}$$

where

$$\begin{aligned} K_{\mathrm{Airy}} &\doteq \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{x-y} \\ &\qquad \text{acting on } L^2(s,\infty) \text{ (Airy kernel)} \end{aligned}$$

and q is the unique solution to the Painlevé II equation

$$q'' = sq + 2q^3$$

satisfying the condition

$$q(s) \sim \mathrm{Ai}(s) \text{ as } s \to \infty.$$

The orthogonal and symplectic distribution functions are

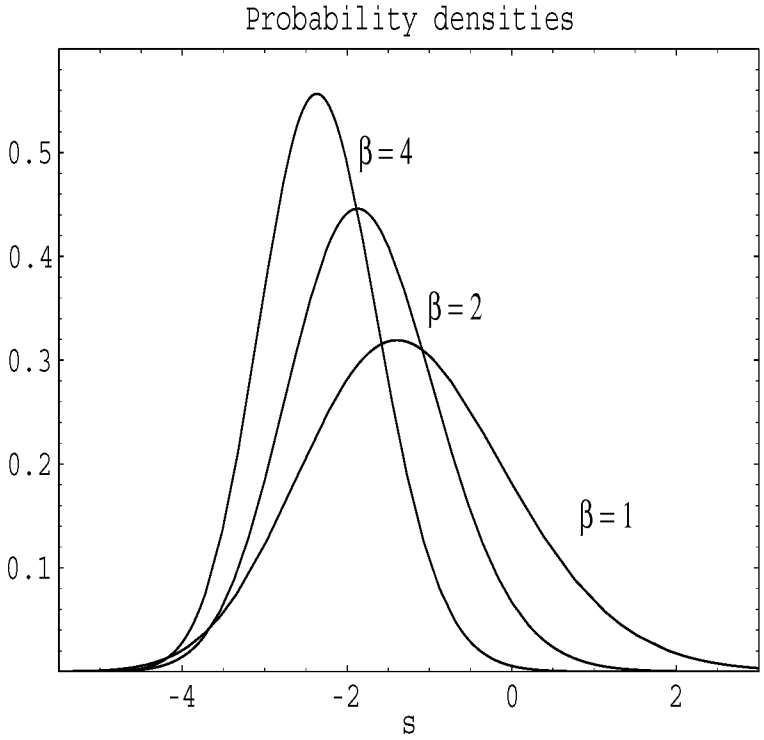
$$\begin{aligned} F_1(s) &= \exp\left(-\frac{1}{2}\int_s^\infty q(x)\,dx\right)\left(F_2(s)\right)^{1/2}, \\ F_4(s/\sqrt{2}) &= \cosh\left(\frac{1}{2}\int_s^\infty q(x)\,dx\right)\left(F_2(s)\right)^{1/2}. \end{aligned}$$

Graphs of the densities dF_β/ds are in the adjacent figure and some statistics of F_β can be found in the table.

[1] Here σ is the standard deviation of the Gaussian distribution on the off-diagonal matrix elements. For the normalization we've chosen, $\sigma = 1/\sqrt{2}$; however, for subsequent comparisons, the normalization $\sigma = \sqrt{N}$ is perhaps more natural.

Table 1: The mean (μ_β), standard deviation (σ_β), skewness (S_β) and kurtosis (K_β) of F_β.

β	μ_β	σ_β	S_β	K_β
1	-1.20653	1.2680	0.293	0.165
2	-1.77109	0.9018	0.224	0.093
4	-2.30688	0.7195	0.166	0.050



The Airy kernel is an example of an *integrable integral operator* [19] and a general theory is developed in [49]. A vertex operator approach to these distributions (and many other closely related distribution functions in random matrix theory) was initiated by Adler, Shiota and van Moerbeke [1]. (See the review article [51] for further developments of this latter approach.)

Historically, the discovery of the connection between Painlevé functions (P_{III} in this case) and Toeplitz/Fredholm determinants appears in work of Wu et al. [53] on the spin-spin correlation functions of the two-dimensional Ising model. Painlevé functions first appear in random matrix theory in Jimbo et al. [20] where they prove the Fredholm determinant of the sine kernel is expressible in terms of P_V.

Gaudin [13] (using Mehta's [30] then newly invented method of orthogonal polynomials) was the first to discover the connection between random matrix theory and Fredholm determinants.

1.1. Universality theorems

A natural question is to ask whether the above limit laws depend upon the underlying Gaussian assumption on the probability measure. To investigate this for unitarily invariant measures ($\beta = 2$) one replaces in (1.1)

$$\exp\left(-\mathrm{tr}(A^2)\right) \to \exp\left(-\mathrm{tr}(V(A))\right).$$

Bleher and Its [9] choose

$$V(A) = gA^4 - A^2, g > 0,$$

and subsequently a large class of potentials V was analyzed by Deift et al. [12]. These analyses require proving new Plancherel-Rotach type formulas for nonclassical orthogonal polynomials. The proofs use Riemann-Hilbert methods. It was shown that the generic behavior is GUE; and hence, the limit law for the largest eigenvalue is F_2. However, by finely tuning the potential new universality classes will emerge at the edge of the spectrum. For $\beta = 1, 4$ a universality theorem was proved by Stojanovic [44] for the quartic potential.

In the case of noninvariant measures, Soshnikov [42] proved that for real symmetric Wigner matrices[2] (complex hermitian Wigner matrices) the limiting distribution of the largest eigenvalue is F_1 (respectively, F_2). The significance of this result is that nongaussian Wigner measures lie outside the "integrable class" (e.g. there are no Fredholm determinant representations for the distribution functions) yet the limit laws are the same as in the integrable cases.

2. Appearance of F_β in limit theorems

In this section we briefly survey the appearances of the limit laws F_β in widely differing areas.

2.1. Combinatorics

A major breakthrough ocurred with the work of Baik, Deift and Johansson [3] when they proved that the limiting distribution of the length of the longest increasing subsequence in a random permutation is F_2. Precisely, if $\ell_N(\sigma)$ is the length of the longest increasing subsequence in the permutation $\sigma \in S_N$, then

$$\mathrm{P}\left(\frac{\ell_N - 2\sqrt{N}}{N^{1/6}} < s\right) \to F_2(s)$$

[2] A symmetric Wigner matrix is a random matrix whose entries on and above the main diagonal are independent and identically distributed random variables with distribution function F. Soshnikov assumes F is even and all moments are finite.

as $N \to \infty$. Here the probability measure on the permutation group S_N is the uniform measure. Further discussion of this result can be found in Johansson's contribution to these proceedings [26].

Baik and Rains [5, 6] showed by restricting the set of permutations (and these restrictions have natural symmetry interpretations) F_1 and F_4 also appear. Even the distributions F_1^2 and F_2^2 [50] arise. By the Robinson-Schensted-Knuth correspondence, the Baik-Deift-Johansson result is equivalent to the limiting distribution on the number of boxes in the first row of random standard Young tableaux. (The measure is the push-forward of the uniform measure on S_N.) These same authors conjectured that the limiting distributions of the number of boxes in the second, third, etc. rows were the same as the limiting distributions of the next-largest, next-next-largest, etc. eigenvalues in GUE. Since these eigenvalue distributions were also found in [47], they were able to compare the then unpublished numerical work of Odlyzko and Rains [34] with the predicted results of random matrix theory. Subsequently, Baik, Deift and Johansson [4] proved the conjecture for the second row. The full conjecture was proved by Okounkov [33] using topological methods and by Johansson [23] and by Borodin, Okounkov and Olshanski [10] using analytical methods. For an interpretation of the Baik-Deift-Johansson result in terms of the card game *patience sorting*, see the very readable review paper by Aldous and Diaconis [2].

2.2. Growth processes

Growth processes have an extensive history both in the probability literature and the physics literature (see, e.g. [15, 29, 41] and references therein), but it was only recently that Johansson [22, 26] proved that the *fluctuations* about the limiting shape in a certain growth model (Corner Growth Model) are F_2. Johansson further pointed out that certain symmetry constraints (inspired from the Baik-Rains work [5, 6]), lead to F_1 fluctuations. This growth model is in Johansson's contribution to these proceedings [26] where the close analogy to largest eigenvalue distributions is explained.

Subsequently, Baik and Rains [7] and Gravner, Tracy and Widom [16] have shown the same distribution functions appearing in closely related lattice growth models. Prähofer and Spohn [38, 39] reinterpreted the work of [3] in terms of the physicists' polynuclear growth model (PNG) thereby clarifying the role of the symmetry parameter β. For example, $\beta = 2$ describes growth from a single droplet where as $\beta = 1$ describes growth from a flat substrate. They also related the distributions functions F_β to fluctuations of the height function in the KPZ equation [28, 29]. (The connection with the KPZ equation is heuristic.) Thus one expects on physical grounds that the fluctuations of any growth process falling into the $1 + 1$ KPZ universality class will be described by the distribution functions F_β or one of the generalizations by Baik and Rains [7]. Such a physical conjecture can be tested experimentally; and indeed, Timonen and his colleagues [45] have taken up this challenge. Earlier Timonen et al. [32] established experimentally that a slow, flameless burning process in a random medium (paper!) is in the $1 + 1$ KPZ universality class. This sequence of events is a rare instance in which new results in mathematics inspires

new experiments in physics.

In the context of the PNG model, Prähofer and Spohn have given a process interpretation, the *Airy process*, of F_2. Further work in this direction can be found in Johansson [25].

There is an extension of the growth model in [16] to growth in a *random environment.* In [17] the following model of interface growth in two dimensions is considered by introducing a height function on the sites of a one-dimensional integer lattice with the following update rule: the height above the site x increases to the height above $x-1$, if the latter height is larger; otherwise the height above x increases by one with probability p_x. It is assumed that the p_x are chosen independently at random with a common distribution function F, and that the intitial state is such that the origin is far above the other sites. In the *pure regime* Gravner-Tracy-Widom identify an asymptotic shape and prove that the fluctuations about that shape, normalized by the square root of the time, are asymptotically normal. This constrasts with the quenched version: conditioned on the environment and normalized by the cube root of time, the fluctuations almost surely approach the distribution function F_2. We mention that these same authors in [18] find, under some conditions on F at the right edge, a *composite regime* where now the interface fluctuations are governed by the extremal statistics of p_x in the annealed case while the fluctuations are asymptotically normal in the quenched case.

2.3. Random tilings

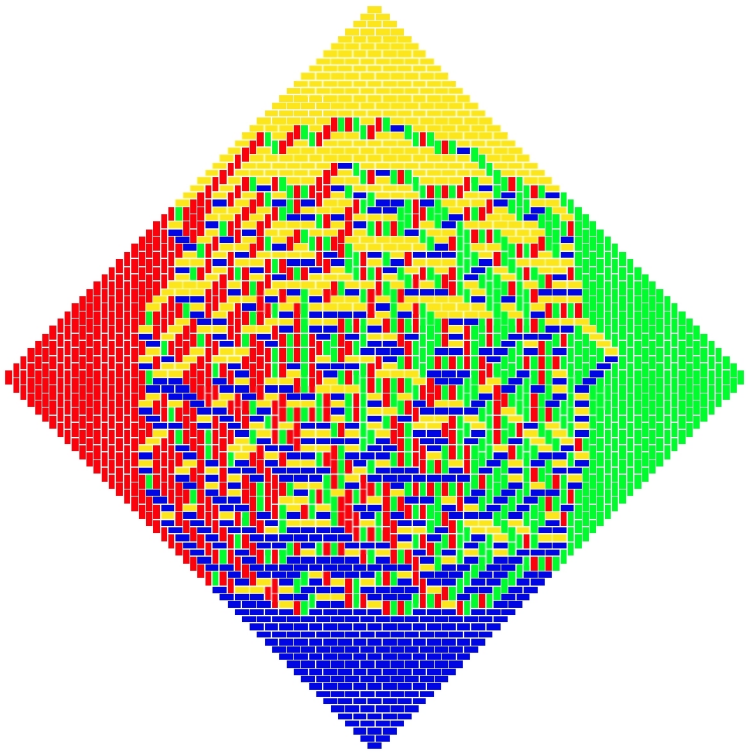
The Aztec diamond of order n is a tiling by dominoes of the lattice squares $[m, m+1] \times [\ell, \ell+1]$, $m, n \in \mathbb{Z}$, that lie inside the region $\{(x,y) : |x| + |y| \leq n+1\}$. A *domino* is a closed 1×2 or 2×1 rectangle in $\mathbb{R}^2$ with corners in $\mathbb{Z}^2$. A typical tiling is shown in the accompanying figure. One observes that near the center the tiling appears random, called the *temperate zone*, whereas near the edges the tiling is frozen, called the *polar zones.* It is a result of Jockush, Propp and Shor [21] (see also [11]) that as $n \to \infty$ the boundary between the temperate zone and the polar zones (appropriately scaled) converges to a circle (Arctic Circle Theorem). Johansson [24] proved that the fluctuations about this limiting circle are F_2.

2.4. Statistics

Johnstone [27] considers the largest principal component of the covariance matrix $X^t X$ where X is an $n \times p$ data matrix all of whose entries are independent standard Gaussian variables and proves that for appropriate centering and scaling, the limiting distribution equals F_1 in the limit $n, p \to \infty$ with $n/p \to \gamma \in \mathbb{R}^+$. Soshnikov [43] has removed the Gaussian assumption but requires that $n - p = \mathrm{O}(p^{1/3})$. Thus we can anticipate applications of the distributions F_β (and particularly F_1) to the statistical analysis of large data sets.

2.5. Queuing theory

Glynn and Whitt [14] consider a series of n single-server queues each with unlimited waiting space with a first-in and first-out service. Service times are i.i.d. with



Random Tilings Research Group

mean one and variance σ^2 with distribution V. The quantity of interest is $D(k,n)$, the departure time of customer k (the last customer to be served) from the last queue n. For a fixed number of customers, k, they prove that

$$\frac{D(k,n)-n}{\sigma\sqrt{n}}$$

converges in distribution to a certain functional $\hat{D}_k$ of k-dimensional Brownian motion. They show that $\hat{D}_k$ is independent of the service time distribution V. It was shown in [8, 16] that $\hat{D}_k$ is equal in distribution to the largest eigenvalue of a $k \times k$ GUE random matrix. This fascinating connection has been greatly clarified in recent work of O'Connell and Yor [35] (see also [36]).

From Johansson [22] it follows for V Poisson that

$$\mathrm{P}\left(\frac{D(\lfloor xn \rfloor, n) - c_1 n}{c_2\, n^{1/3}} < s\right) \to F_2(s)$$

as $n \to \infty$ for some explicitly known constants c_1 and c_2 (depending upon x).

2.6. Superconductors

Vavilov et al. [52] have conjectured (based upon certain physical assumptions supported by numerical work) that the fluctuation of the excitation gap in a metal grain or quantum dot induced by the proximity to a superconductor is described by F_1 for zero magnetic field and by F_2 for nonzero magnetic field. They conclude their paper with the remark:

> The universality of our prediction should offer ample opportunities for experimental observation.

Acknowledgements: This work was supported by the National Science Foundation through grants DMS-9802122 and DMS-9732687.

References

[1] M. Adler, T. Shiota and P. van Moerbeke, Random matrices, vertex operators and the Virasoro algebra, *Phys. Letts. A* **208** (1995), 67–78.

[2] D. Aldous and P. Diaconis, Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theory, *Bull. Amer. Math. Soc.* **36** (1999), 413–432.

[3] J. Baik, P. Deift and K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations, *J. Amer. Math. Soc.* **12** (1999), 1119–1178.

[4] J. Baik, P. Deift and K. Johansson, On the distribution of the length of the second row of a Young diagram under Plancherel measure, *Geom. Funct. Anal.* **10** (2000), 702–731.

[5] J. Baik and E. M. Rains, The asymptotics of monotone subsequences of involutions, *Duke Math. J.* **109** (2001), 205–281.

[6] J. Baik and E. M. Rains, Symmetrized random permutations, in *Random Matrix Models and their Applications*, eds. P. Bleher and A. Its, Math. Sci. Res. Inst. Publications **40**, Cambridge Univ. Press, 2001, 1–19.

[7] J. Baik and E. M. Rains, Limiting distributions for a polynuclear growth model, *J. Stat. Phys.* **100** (2000), 523–541.

[8] Yu. Baryshnikov, GUEs and queues, *Probab. Th. Rel. Fields* **119** (2001), 256–274.

[9] P. Bleher and A. Its, Semiclassical asymptotics of orthongonal polynomials, Riemann-Hilbert problem, and universality in the matrix model, *Ann. Math.* **150** (1999), 185–266.

[10] A. Borodin, A. Okounkov and G. Olshanski, Asymptotics of Plancherel measures for symmetric groups, *J. Amer. Math. Soc.* **13** (2000), 481–515.

[11] H. Cohn, N. Elkies and J. Propp, Local statistics for random domino tilings of the Aztec diamond, *Duke Math. J.* **85** (1996), 117–166.

[12] P. Deift, T. Kriecherbauer, K. T-R. McLauglin, S. Venakides and X. Zhou, Uniform asymptotics for polynomials orthogonal with respect to varying exponential weight and applications to universality questions in random matrix theory, *Commun. Pure Appl. Math.* **52** (1999), 1335–1425.

[13] M. Gaudin, Sur la dlo limite de l'espacement des valeurs propres d'une matrice aléatoire, *Nucl. Phys.* **25** (1961), 447–458; reprinted in [37].
[14] P. W. Glynn and W. Whitt, Departure times from many queues, *Ann. Appl. Probab.* **1** (1991), 546–572.
[15] J. Gravner and D. Griffeath, Cellular automaton growth on $\mathbb{Z}^2$: Theorems, examples and problems, *Adv. Appl. Math.* **21** (1998), 241–304.
[16] J. Gravner, C. A. Tracy and H. Widom, Limit theorems for height fluctuations in a class of discrete space and time growth models, *J. Stat. Phys.* **102** (2001), 1085–1132.
[17] J. Gravner, C. A. Tracy and H. Widom, A growth model in a random environment, *Ann. Probab.* **30** (2002), 1340–1368.
[18] J. Gravner, C. A. Tracy and H. Widom, Fluctuations in the composite regime of a disordered growth model, *Commun. Math. Phys.* **229** (2002), 433–458.
[19] A. R. Its, A. G. Izergin, V. E. Korepin and N. A. Slavnov, Differential equations for quantum correlations, *J. Mod. Phys. B* **4** (1990), 1003–1037.
[20] M. Jimbo, T. Miwa, Y. Môri and M. Sato, Density matrix of an impenetrable Bose gas and the fifth Painlevé transcendent, *Physica D* **1** (1980), 80–158.
[21] W. Jockush, J. Propp and P. Shor, Random domino tilings and the arctic circle theorem, preprint.
[22] K. Johansson, Shape fluctuations and random matrices, *Commun. Math. Phys.* **209** (2000), 437–476.
[23] K. Johansson, Discrete orthogonal polynomial ensembles and the Plancherel measure, *Ann. Math.* **153** (2001), 259–296.
[24] K. Johansson, Non-intersecting paths, random tilings and random matrices, *Probab. Th. Rel. Fields 123* (2002), 225–280.
[25] K. Johansson, Discrete polynuclear growth and determinantal processes, preprint (arXiv: math.PR/0206208).
[26] K. Johansson, Toeplitz determinants, random growth and determinantal processes, ICM 2002, Vol. III, 53–62.
[27] I. Johnstone, On the distribution of the largest principal component, *Ann. Statistics* **29** (2001), 295–327.
[28] M. Kardar, G. Parisi and Y-C Zhang, Dynamical scaling of growing interfaces, *Phys. Rev. Letts.* **56** (1986), 889–892.
[29] P. Meakin, *Fractals, Scaling and Growth Far from Equilibrium*, Cambridge Univ. Press, 1998.
[30] M. L. Mehta, On the statistical properties of the level-spacings in nuclear spectra, *J. Math. Phys.* **18** (1960), 395–419; reprinted in [37].
[31] M. L. Mehta, *Random Matrices*, second ed., Academic Press, 1991.
[32] M. Myllys, J. Maunuksela, M. Alava, T. Ala-Nissila, J. Merikoski and J. Timonen, Kinetic roughening in slow combustion of paper, *Phys. Rev. E* **64**, 036101-1–036101-12.
[33] A. Okounkov, Random matrices and random permutations, *Internat. Math. Res. Notices* no. 20 (2000), 1043–1095.
[34] A. M. Odlyzko and E. M. Rains, On the longest increasing subsequences in random permutations, in *Analysis, Geometry, Number Theory: The Mathematics*

of Leon Ehrenpreis, eds. E. L. Grinberg, S. Berhanu, M. Knopp, G. Mendoza and E. T. Quinto, Amer. Math. Soc. 2000, 439–451.

[35] N. O'Connell and M. Yor, A representation for non-colliding random walks, *Elect. Commun. in Probab.* **7** (2002), 1–12.

[36] N. O'Connell, Random matrices, non-colliding processes and queues, preprint (arXiv: math/PR/0203176).

[37] C. E. Porter, *Statistical Theories of Spectra: Fluctuations*, Academic Press, 1965.

[38] M. Prähofer and H. Spohn, Statistical self-similarity of one-dimensional growth processes, *Physica A* **279** (2000), 342–352.

[39] M. Prähofer and H. Spohn, Universal distributions for growth processes in 1+1 dimensions and random matrices, *Phys. Rev. Letts.* **84** (2000), 4882–4885.

[40] M. Prähofer and H. Spohn, Scale invariance of the PNG droplet and the Airy process, J. Statistical Physcics **108** (2002), 1071–1106.

[41] T. Seppäläinen, Exact limiting shape for a simplified model of first-passage percolation in the plane, *Ann. Prob.* **26** (1998), 1232–1250.

[42] A. Soshnikov, Universality at the edge of the spectrum in Wigner random matrices, *Commun. Math. Phys.* **207** (1999), 697–733.

[43] A. Soshnikov, A note on universality of the distribution of the largest eigenvalue in certain classes of sample covariance matrices, preprint (arXiv: math.PR/0104113).

[44] A. Stojanovic, Universality in orthogonal and symplectic invariant matrix models with quartic potential, *Math. Phys., Analy. and Geom.* **3** (2000), 339–373.

[45] J. Timonen, private communication.

[46] C. A. Tracy and H. Widom, Level-spacing distribution and the Airy kernel, *Phys. Letts. B* **305** (1993), 115–118.

[47] C. A. Tracy and H. Widom, Level-spacing distribution and the Airy kernel, *Commun. Math. Phys.* **159** (1994), 151–174.

[48] C. A. Tracy and H. Widom, On orthogonal and symplectic matrix ensembles, *Commun. Math. Phys.* **177** (1996), 727–754.

[49] C. A. Tracy and H. Widom, Fredholm determinants, differential equations and matrix models, *Commun. Math. Phys.* **163** (1994), 151–174.

[50] C. A. Tracy and H. Widom, Random unitary matrices, permutations and Painlevé, *Commun. Math. Phys.* **207** (1999), 665–685.

[51] P. van Moerbeke, Integrable lattices: Random matrices and random permutations, in *Random Matrix Models and their Applications*, eds. P. Bleher and A. Its, Math. Sci. Res. Inst. Publications **40**, Cambridge Univ. Press, 2001, 321–406.

[52] M. G. Vavilov, P. W. Brouwer, V. Ambegaokar and C. W. J. Beenakker, Universal gap fluctuations in the superconductor proximity effect, *Phys. Rev. Letts.* **86** (2001), 874–877.

[53] T. T. Wu, B. M. McCoy, C. A. Tracy and E. Barouch, Spin-spin correlation functions of the two-dimensional Ising model: Exact theory in the scaling region, *Phys. Rev. B* **13** (1976), 316–374.

ICM 2002 · Vol. I · 597–606

Smoothed Analysis of Algorithms

Daniel A. Spielman* Shang-Hua Teng†

Abstract

Spielman and Teng [STOC '01] introduced the smoothed analysis of algorithms to provide a framework in which one could explain the success in practice of algorithms and heuristics that could not be understood through the traditional worst-case and average-case analyses. In this talk, we survey some of the smoothed analyses that have been performed.

2000 Mathematics Subject Classification: 65Y20, 68Q17, 68Q25, 90C05.
Keywords and Phrases: Smoothed analysis, Simplex method, Condition number, Interior point method, Perceptron algorithm.

1. Introduction

The most common theoretical approach to understanding the behavior of algorithms is worst-case analysis. In worst-case analysis, one proves a bound on the worst possible performance an algorithm can have. A triumph of the Algorithms community has been the proof that many algorithms have good worst-case performance—a strong guarantee that is desirable in many applications. However, there are many algorithms that work exceedingly well in practice, but which are known to perform poorly in the worst-case or lack good worst-case analyses. In an attempt to rectify this discrepancy between theoretical analysis and observed performance, researchers introduced the average-case analysis of algorithms. In average-case analysis, one bounds the expected performance of an algorithm on random inputs. While a proof of good average-case performance provides evidence that an algorithm may perform well in practice, it can rarely be understood to explain the good behavior of an algorithm in practice. A bound on the performance of an algorithm under one distribution says little about its performance under another distribution, and may say little about the inputs that occur in practice. Smoothed analysis is a hybrid of worst-case and average-case analyses that inherits advantages of both.

*Department of Mathematics, MIT, USA. E-mail: spielman@math.mit.edu
†Department of Computer Science, Boston University, USA.

In the formulation of smoothed analysis used in [27], we measure the maximum over inputs of the expected running time of a simplex algorithm under slight random perturbations of those inputs. To see how this measure compares with worst-case and average-case analysis, let X_n denote the space of linear-programming problems of length n and let $\mathcal{T}(x)$ denote the running time of the simplex algorithm on input x. Then, the worst-case complexity of the simplex algorithm is the function

$$\mathcal{C}_{worst}(n) = \max_{x \in X_n} \mathcal{T}(x),$$

and the average-case complexity of the algorithm is

$$\mathcal{C}_{ave}(n) = \mathbf{E}_{r \in X_n} \mathcal{T}(r),$$

under some suitable distribution on X_n. In contrast, the smoothed complexity of the simplex algorithm is the function

$$\mathcal{C}_{smooth}(n, \sigma) = \max_{x} \mathbf{E}_{r \in X_n} \mathcal{T}(x + \sigma \|x\| r),$$

where r is chosen according to some distribution, such as a Gaussian. In this case $\sigma \|x\| r$ is a Gaussian random vector of standard deviation $\sigma \|x\|$. We multiply by $\|x\|$ so that we can relate the magnitude of the perturbation to the magnitude of that which it perturbs.

In the smoothed analysis of algorithms, we measure the expected performance of algorithms under slight random perturbations of worst-case inputs. More formally, we consider the maximum over inputs of the expected performance of algorithms under slight random perturbations of those inputs. We then express this expectation as a function of the input size and the magnitude of the perturbation. While an algorithm with a good worst-case analysis will perform well on all inputs, an algorithm with a good smoothed analysis will perform well on almost all inputs in every small neighborhood of inputs. Smoothed analysis makes sense for algorithms whose inputs are subject to slight amounts of noise in their low-order digits, which is typically the case if they are derived from measurements of real-world phenomena. If an algorithm takes such inputs and has a good smoothed analysis, then it is unlikely that it will encounter an input on which it performs poorly. The name "smoothed analysis" comes from the observation that if one considers the running time of an algorithm as a function from inputs to time, then the smoothed complexity of the algorithm is the highest peak in the plot of this function after it is convolved with a small Gaussian.

In our paper introducing smoothed analysis, we proved that the simplex method has polynomial smoothed complexity [27]. The simplex method, which has been the most popular method of solving linear programs since the late 1940's, is the canonical example of a practically useful algorithm that could not be understood theoretically. While it was known to work very well in practice, contrived examples on which it performed poorly proved that it had horrible worst-case complexity[19, 20, 14, 13, 4, 17, 3]. The average-case complexity of the simplex method was proved to be polynomial [8, 7, 25, 16, 1, 2, 28], but this result was not considered to explain the performance of the algorithm in practice.

2. The simplex method for linear programming

We recall that a linear programming problem can be written in the form

$$\begin{aligned} &\text{maximize} && x^T c \\ &\text{subject to} && x^T a_i \leq b_i, \text{ for } 1 \leq i \leq n, \end{aligned} \tag{2.1}$$

where $c \in \mathbb{R}^d$, $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$, for $1 \leq i \leq n$ In [27], we bound the smoothed complexity of a particular two-phase simplex method that uses the shadow-vertex pivot rule to solve linear programs in this form.

We recall that the constraints of the linear program, that $x^T a_i \leq b_i$, confine x to a (possibly open) polytope, and that the solution to the linear program is a vertex of this polytope. Simplex methods work by first finding some vertex of the polytope, and then walking along the 1-faces of the polytope from vertex to vertex, improving the objective function at each step. The pivot rule of a simplex algorithm dictates which vertex the algorithm should walk to when it has many to choose from. The shadow-vertex method is inspired by the simplicity of the simplex method in two-dimensions: in two-dimensions, the polytope is a polygon and the choice of next vertex is always unique. To lift this simplicity to higher dimensions, the shadow-vertex simplex method considers the orthogonal projection of the polytope defined by the constraints onto a two-dimensional space. The method then walks along the vertices of the polytope that are the pre-images of the vertices of the shadow polygon. By taking the appropriate shadow, it is possible to guarantee that the vertex optimizing the objective function will be encountered during this walk. Thus, the running time of the algorithm may be bounded by the number of vertices lying on the shadow polygon. Our first step in proving a bound on this number is a smoothed analysis of the number of vertices in a shadow. For example, we prove the bound:

Theorem 2.1 (Shadow Size) *Let $d \geq 3$ and $n > d$. Let c and t be independent vectors in $\mathbb{R}^d$, and let $a_1, \ldots, a_n$ be Gaussian random vectors in $\mathbb{R}^d$ of variance $\sigma^2 \leq \frac{1}{9d \log n}$ centered at points each of norm at most 1. Then, the expected number of vertices of the shadow polygon formed by the projection of $\{x : x^T a_i \leq 1\}$ onto* $\mathbf{Span}(t, c)$ *is at most*

$$\frac{58,888,678\ nd^3}{\sigma^6}. \tag{2.2}$$

This bound does not immediately lead to a bound on the running time of a shadow-vertex method as it assumes that t and c are fixed before the a_is are chosen, while in a simplex method the plane on which the shadow is followed depends upon the a_is. However, we are able to use the shadow size bound as a black-box to prove for a particular randomized two-phase shadow vertex simplex method:

Theorem 2.2 (Simplex Method) *Let $d \geq 3$ and $n \geq d + 1$. Let $c \in \mathbb{R}^d$ and $b \in \{-1, 1\}^n$. Let $a_1, \ldots, a_n$ be Gaussian random vectors in $\mathbb{R}^d$ of variance $\sigma^2 \leq \frac{1}{9d \log n}$ centered at points each of norm at most 1. Then the expected number of simplex steps taken by the two-phase shadow-vertex simplex algorithm to solve the program specified by b, c, and $a_1, \ldots, a_n$ is at most*

$$(nd/\sigma)^{O(1)},$$

where the expectation is over the choice of $a_1, \ldots, a_n$ and the random choices made by the algorithm.

While the proofs of Theorems 2.1 and 2.2 are quite involved, we can provide the reader with this intuition for Theorem 2.1: *after perturbation, most of the vertices of the polytope defined by the linear program have an angle bounded away from flat.* This statement is not completely precise because "most" should be interpreted under a measure related to the chance a vertex appears in the shadow, as opposed to counting the number of vertices. Also, there are many ways of measuring high-dimensional angles, and different approaches are used in different parts of the proof. However, this intuitive statement tells us that most vertices on the shadow polygon should have angle bounded away from flat, which means that there cannot be too many of them.

One way in which angles of vertices are measured is by the condition number of their defining equations. A vertex of the polytope is given by a set of equations of the form

$$Cx = b.$$

The condition number of C is defined to be

$$\kappa(C) = \|C\| \left\|C^{-1}\right\|,$$

where we recall that

$$\|C\| = \max_x \frac{\|Cx\|}{\|x\|}$$

and that

$$\left\|C^{-1}\right\| = \min_x \frac{\|Cx\|}{\|x\|}.$$

The condition number is a measure of the sensitivity of x to changes in C and b, and is also a normalized measure of the distance of C to the set of singular matrices. For more information on the condition number of a matrix, we refer the reader to one of [15, 29, 10]. Condition numbers play a fundamental role in Numerical Analysis, which we will now discuss.

3. Smoothed complexity framework for numerical analysis

The condition number of a problem instance is generally defined to be the sensitivity of the output to slight perturbations of the problem instance. In Numerical Analysis, one often bounds the running time of an iterative algorithm in terms of the condition number of its input. Classical examples of algorithms subject to such analyses include Newton's method for root finding and the conjugate gradient method of solving systems of linear equations. For example, the number of iterations taken by the method of conjugate gradients is proportional to the square root of the condition number. Similarly, the running times of interior-point methods have been bounded in terms of condition numbers [22].

Blum [6] suggested that a complexity theory of numerical algorithms should be parameterized by the condition number of an input in addition to the input size. Smale [26] proposed a complexity theory of numerical algorithms in which one:

1. *proves a bound on the running time of an algorithm solving a problem in terms of its condition number, and then*
2. *proves that it is unlikely that a random problem instance has large condition number.*

This program is analogous to the average-case complexity of Theoretical Computer Science and hence shares the same shortcoming in modeling practical performance of numerical algorithms.

To better model the inputs that occur in practice, we propose replacing step 2 of Smale's program with

2'. *prove that for every input instance it is unlikely that a slight random perturbation of that instance has large condition number.*

That is, we propose to bound the smoothed value of the condition number. In contrast with the average-case analysis of condition numbers, our analysis can be interpreted as demonstrating that if there is a little bit of imprecision or noise in the input, then it is unlikely it is ill-conditioned. The combination of step 2' with step 1 of Smale's program provides a simple framework for performing smoothed analysis of numerical algorithms whose running time can be bounded by the condition number of the input.

4. Condition numbers of matrices

One of the most fundamental condition numbers is the condition number of matrices defined at the end of Section 2. In his paper, "The probability that a numerical analysis problem is difficult", Demmel [9] proved that it is unlikely that a Gaussian random matrix centered at the origin has large condition number. Demmel's bounds on the condition number were improved by Edelman [12]. As bounds on the norm of a random matrix are standard, we focus on the norm of the inverse, for which Edelman proved:

Theorem 4.1 (Edelman) *Let G be a d-by-d matrix of independent Gaussian random variables of variance* 1 *and mean* 0*. Then,*

$$\mathbf{Pr}\left[\left\|G^{-1}\right\| > t\right] \leq \frac{\sqrt{d}}{t}.$$

We obtain a smoothed analogue of this bound in work with Sankar [24]. That is, we show that for every matrix it is unlikely that the slight perturbation of that matrix has large condition number. The key technical statement is:

Theorem 4.2 (Sankar-Spielman-Teng) *Let $\bar{A}$ be an arbitrary d-by-d Real matrix and A a matrix of independent Gaussian random variables centered at $\bar{A}$, each of variance σ^2. Then*

$$\mathbf{Pr}\left[\left\|A^{-1}\right\| > x\right] < 1.823\frac{\sqrt{d}}{x\sigma}$$

In contrast with the techniques used by Demmel and Edelman, the techniques used in the proof of Theorem 4.2 are geometric and completely elementary. We now give the reader a taste of these techniques by proving the simpler:

Theorem 4.3 *Let $\bar{A}$ be an arbitrary d-by-d Real matrix and A a matrix of independent Gaussian random variables centered at $\bar{A}$, each of variance σ^2. Then,*

$$\mathbf{Pr}\left[\left\|A^{-1}\right\| \geq x\right] \leq d^{3/2}/x\sigma.$$

The first step of the proof is to relate $\left\|A^{-1}\right\|$ to a geometric quantity of the vectors in the matrix A. The second step is to bound the probability of a configuration under which this geometric quantity is small. The geometric quantity is given by:

Definition *For d vectors in $\mathbb{R}^d$, $a_1, \ldots, a_d$, define*

$$\mathbf{height}\,(a_1, \ldots, a_d) = \min_i \mathbf{dist}\,(a_i, \mathbf{Span}\,(a_1, \ldots, \hat{a_i}, \ldots, a_d)).$$

Lemma 4.5 *For d vectors in $\mathbb{R}^d$, $a_1, \ldots, a_d$,*

$$\left\|(a_1, \ldots, a_d)^{-1}\right\| \leq \sqrt{d}/\mathbf{height}\,(A).$$

Proof. Let t be a unit vector such that

$$\left\|\sum_{i=1}^{d} t_i a_i\right\| = 1/\left\|(a_1, \ldots, a_d)^{-1}\right\|.$$

Without loss of generality, let t_1 be the largest entry of t in absolute value, so $|t_1| \geq 1/\sqrt{d}$. Then, we have

$$\begin{aligned} \left\|a_1 + \sum_{i=2}^{d}(t_i/t_1)a_i\right\| &\leq \sqrt{n}/\left\|(a_1, \ldots, a_d)^{-1}\right\| \\ \Longrightarrow \mathbf{dist}\,(a_1, \mathbf{Span}\,(a_2, \ldots, a_d)) &\leq \sqrt{d}/\left\|(a_1, \ldots, a_d)^{-1}\right\|. \quad \spadesuit \end{aligned}$$

Proof of Theorem 4.3. Let $a_1, \ldots, a_d$ denote the columns of A. Lemma 4.5 tells us that if $\left\|A^{-1}\right\| \geq x$, then $\mathbf{height}\,(a_1, \ldots, a_d)$ is less than $\sqrt{d}/x$. For each i, the probability that the height of a_i above $\mathbf{Span}\,(a_1, \ldots, \hat{a_i}, \ldots, a_d)$ is less than $\sqrt{d}/x$ is at most

$$\sqrt{d}/x\sigma.$$

Thus,

$$\begin{aligned} &\mathbf{Pr}\left[\mathbf{height}\,(a_1, \ldots, a_d) < \sqrt{d}/x\right] \\ &\leq \mathbf{Pr}\left[\exists i : \mathbf{dist}\,(a_i, \mathbf{Span}\,(a_1, \ldots, \hat{a_i}, \ldots, a_n)) < \sqrt{d}/x\right] \\ &< n^{3/2}/x\sigma; \end{aligned}$$

so,

$$\mathbf{Pr}\left[\left\|(a_1, \ldots, a_d)^{-1}\right\| > x\right] < d^{3/2}/x\sigma. \quad \spadesuit$$

Conjecture 1 *Let $\bar{A}$ be an arbitrary d-by-d Real matrix and A a matrix of independent Gaussian random variables centered at $\bar{A}$, each of variance σ^2. Then*

$$\mathbf{Pr}\left[\left\|A^{-1}\right\| > x\right] < \frac{\sqrt{d}}{x\sigma}.$$

5. Smoothed condition numbers of linear programs

The Perceptron algorithm solves linear programs of the following simple form:

> Given a set of points $a_1, \ldots, a_n$, find a vector x such that $\langle a_i | x \rangle > 0$ for all i, if one exists.

One can define the condition number of the Perceptron problem to be the reciprocal of the "wiggle room" of the input. That is, let $S = \{x : \langle a_i | x \rangle > 0, \forall i\}$ and

$$\nu(a_1, \ldots, a_n) = \max_{x \in S} \left(\min_i \frac{|\langle a_i | x \rangle|}{\|a_i\| \, \|x\|} \right).$$

Then, the condition number of Perceptron problem is defined to be $1/\nu(a_1, \ldots, a_n)$.

The Perceptron algorithm works as follows: (1) Initialize $x = \mathbf{0}$; (2) Select any a_i such that $\langle a_i | x \rangle \leq 0$ and set $x = x + a_i / \|a_i\|$; (3) while $x \notin S$, go back to step (2).

Using the following two lemmas, Blum and Dunagan [5] obtained a smoothed analysis of the Perceptron algorithm.

Theorem 5.1 (Block-Novikoff) *On input $a_1, \ldots, a_n$, the perceptron algorithm terminates in at most $1/(\nu(a_1, \ldots, a_n))^2$ iterations.*

Theorem 5.2 (Blum-Dunagan) *Let $a_1, \ldots, a_n$ be Gaussian random vectors in $\mathbb{R}^d$ of variance $\sigma^2 < 1/(2d)$ centered at points each of norm at most 1. Then,*

$$\mathbf{Pr}\left[\frac{1}{\nu(a_1, \ldots, a_n)} > t\right] \leq \frac{nd^{1.5}}{\sigma t} \log \frac{\sigma t}{d^{1.5}}.$$

Setting $t = \frac{nd^{1.5} \log(n/\delta)}{\delta \sigma}$, Blum and Dunagan concluded

Theorem 5.3 (Blum-Dunagan) *Let $a_1, \ldots, a_n$ be Gaussian random vectors in $\mathbb{R}^d$ of variance $\sigma^2 < 1/(2d)$ centered at points each of norm at most 1. Then, there exists a constant c such that the probability that the perceptron takes more than $\frac{cd^3 n^2 \log^2(n/\delta)}{\delta^2 \sigma^2}$ iterations is at most δ.*

In his seminal work, Renegar [21, 22, 23] defines the condition of a linear program to be the normalized reciprocal of its distance to the set of ill-posed linear programs, where an ill-posed program is one that can be made both feasible and infeasible or bounded and unbounded by arbitrarily small changes to its constraints.

Renegar proved the following theorem.

Theorem 5.4 (Renegar) *There is an interior point method such that, on input a linear program specified by (A, b, c) and an $\epsilon > 0$, it will terminate in*

$$O(\sqrt{n+d} \log(\kappa(A, b, c)/\epsilon)$$

iterations and return either a solution within ϵ of the optimal or a certificate that the linear program is infeasible or unbounded.

With Dunagan, we recently proved the following smoothed bound on the condition number of a linear program [11]:

Theorem 5.5 (Dunagan-Spielman-Teng) *For any $\sigma^2 < 1/(nd)$, let $A = (a_1, \ldots, a_n)$ be a set of Gaussian random vectors in $\mathbb{R}^d$ of variance σ^2 centered at points $\bar{a}_1, \ldots, \bar{a}_n$, let b be a Gaussian random vector in $\mathbb{R}^d$ of variance σ^2 centered at $\bar{b}$ and let c be a Gaussian random vector in $\mathbb{R}^n$ of variance σ^2 centered at $\bar{c}$ such that $\sum_{i=1}^n \|\bar{a}_i\|^2 + \|\bar{b}\|^2 + \|\bar{c}\|^2 \leq 1$. Then*

$$\mathbf{Pr}_{A,b,c}\left[C(A,b,c) > t\right] < \frac{2^{14} n^2 d^{3/2}}{\sigma^2 t} \log^2 \frac{2^{10} n^2 d^{3/2} t}{\sigma^2},$$

and hence

$$\mathbf{E}_{A,b,c}\left[\log C(A,b,c)\right] \leq 21 + 3\log(nd/\sigma).$$

Combining these two theorem, we have

Theorem 5.6 (Smoothed Complexity of Interior Point Methods) *Let σ and (A,b,c) be as given in Theorem 5.5, Then, Renegar's interior point method solves the linear program specified by (A,b,c) to within precision ϵ in expected*

$$O\left(\sqrt{n+d}(21 + 3\log(nd/\sigma\epsilon))\right)$$

iterations.

6. Two open problems

As the norm of the inverse of matrix is such a fundamental quantity, it is natural to ask how the norms of the inverses of the $\binom{n}{d}$ d-by-d square sub-matrices of a d-by-n matrix behave. Moreover, a crude bound on the probability that many of these are large is a dominant term in the analysis of complexity of the simplex method in [27]. The bound obtained in that paper is:

Lemma 6.1 *Let $a_1, \ldots, a_n$ be Gaussian random vectors in $\mathbb{R}^d$ of variance $\sigma^2 \leq 1/9d\log n$ centered at points of norm at most 1. For $I \in \binom{[n]}{d}$ a d-set, let X_I denote the indicator random variable that is 1 if*

$$\left\| [a_i : i \in I]^{-1} \right\| \geq \frac{\sigma^2}{8 d^{3/2} n^7}.$$

Then,

$$\mathbf{Pr}_{a_1,\ldots,a_n}\left[\frac{d}{2}\sum_I X_I < \left\lceil \frac{n-d-1}{2} \right\rceil \binom{n}{d-1}\right] \geq 1 - n^{-d} - n^{-n+d-1} - n^{-2.9d+1}.$$

Clearly, one should be able to prove a much stronger bound than that stated here, and thereby improve the bounds on the smoothed complexity of the simplex method.

While much is known about the condition numbers of random matrices drawn from continuous distributions, much less is known about matrices drawn from discrete distributions. We conjecture:

Conjecture 2 *Let A be a d-by-d matrix of independently and uniformly chosen ± 1 entries. Then,*

$$\mathbf{Pr}\left[\left\|A^{-1}\right\| > t\right] \leq \frac{\sqrt{d}}{t} + \alpha^n,$$

for some absolute constant $\alpha < 1$.

We remark that the case $t = \infty$, when the matrix A is singular, follows from a theorem of Kahn, Komlos and Szemeredi [18].

References

[1] I. Adler, R. M. Karp, and R. Shamir. A simplex variant solving an m x d linear program in $O(min(m^2, d^2))$ expected number of pivot steps. *J. Complexity*, 3:372–387, 1987.

[2] Ilan Adler and Nimrod Megiddo. A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension. *Journal of the ACM*, 32(4):871–895, October 1985.

[3] Nina Amenta and Gunter Ziegler. Deformed products and maximal shadows of polytopes. In B. Chazelle, J.E. Goodman, and R. Pollack, editors, *Advances in Discrete and Computational Geometry*, number 223 in Contemporary Mathematics, 57–90. Amer. Math. Soc., 1999.

[4] David Avis and Vasek Chvátal. Notes on Bland's pivoting rule. In *Polyhedral Combinatorics*, volume 8 of *Math. Programming Study*, 24–34. 1978.

[5] Avrim Blum and John Dunagan. Smoothed analysis of the perceptron algorithm for linear programming. In *SODA '02*, 905–914, 2002.

[6] Lenore Blum. Lectures on a theory of computation and complexity over the reals (or an arbitrary ring). In Erica Jen, editor, *The Proceedings of the 1989 Complex Systems Summer School, Santa Fe, New Mexico*, volume 2, 1–47, June 1989.

[7] Karl Heinz Borgwardt. *Untersuchungen zur Asymptotik der mittleren Schrittzahl von Simplexverfahren in der linearen Optimierung.* PhD thesis, Universitat Kaiserslautern, 1977.

[8] Karl Heinz Borgwardt. *The Simplex Method: a probabilistic analysis.* Number 1 in Algorithms and Combinatorics. Springer-Verlag, 1980.

[9] James Demmel. The probability that a numerical analysis problem is difficult. *Math. Comp.*, 499–480, 1988.

[10] James Demmel. *Applied Numerical Linear Algebra.* SIAM, 1997.

[11] John Dunagan, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of Renegar's condition number for linear programming. Available at `http://math.mit.edu/~spielman/SmoothedAnalysis`, 2002.

[12] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560, 1988.

[13] Donald Goldfarb. Worst case complexity of the shadow vertex simplex algorithm. Technical report, Columbia University, 1983.

[14] Donald Goldfarb and William T. Sit. Worst case behaviour of the steepest edge simplex method. *Discrete Applied Math*, 1:277–285, 1979.

[15] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins Series in the Mathematical Sciences. The Johns Hopkins University Press and North Oxford Academic, Baltimore, MD, USA and Oxford, England, 1983.

[16] M. Haimovich. The simplex algorithm is very good ! : On the expected number of pivot steps and related properties of random linear programs. Technical report, Columbia University, April 1983.

[17] Robert G. Jeroslow. The simplex algorithm with the pivot rule of maximizing improvement criterion. *Discrete Math.*, 4:367–377, 1973.

[18] Jeff Kahn, Janos Komlos, and Endre Szemeredi. On the probability that a random +/- 1 matrix is singular. *Journal of the American Mathematical Society*, 1995.

[19] V. Klee and G. J. Minty. How good is the simplex algorithm? In Shisha, O., editor, *Inequalities – III*, 159–175. Academic Press, 1972.

[20] K. G. Murty. Computational complexity of parametric linear programming. *Math. Programming*, 19:213–219, 1980.

[21] J. Renegar. Some perturbation theory for linear programming. *Math. Programming*, 65(1, Ser. A):73–91, 1994.

[22] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM J. Optim.*, 5(3):506–524, 1995.

[23] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Math. Programming*, 70(3, Ser. A):279–351, 1995.

[24] Arvind Sankar, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of the condition numbers and growth factors of matrices. Available at `http://math.mit.edu/~spielman/SmoothedAnalysis`, 2002.

[25] S. Smale. On the average number of steps in the simplex method of linear programming. *Mathematical Programming*, 27:241–262, 1983.

[26] Steve Smale. Complexity theory and numerical analysis. *Acta Numerica*, 523–551, 1997.

[27] Daniel Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing (STOC '01)*, 296–305, 2001. Full version available at `http://math.mit.edu/~spielman/SmoothedAnalysis`.

[28] M.J. Todd. Polynomial expected behavior of a pivoting algorithm for linear complementarity and linear programming problems. *Mathematical Programming*, 35:173–192, 1986.

[29] L. N. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, Philadelphia, PA, 1997.

ICM 2002 · Vol. I · 607–620

Adaptive Methods for PDE's Wavelets or Mesh Refinement?

Albert Cohen*

Abstract

Adaptive mesh refinement techniques are nowadays an established and powerful tool for the numerical discretization of PDE's. In recent years, wavelet bases have been proposed as an alternative to these techniques. The main motivation for the use of such bases in this context is their good performances in data compression and the approximation theoretic foundations which allow to analyze and optimize these performances. We shall discuss these theoretical foundations, as well as one of the approaches which has been followed in developing efficient adaptive wavelet solvers. We shall also discuss the similarities and differences between wavelet methods and adaptive mesh refinement.

2000 Mathematics Subject Classification: 65N50, 41A25, 41A46, 42C40.
Keywords and Phrases: Adaptivity, Mesh refinement, Wavelets, Multiscale, Methods, Nonlinear approxmimation.

1. Introduction

Among those relevant phenomenons which are modelled by partial differential or integral equations, countless are the instances where the mathematical solutions exhibit *singularities.* Perhaps the most classical examples are elliptic equations on domains with re-entrant corners, or nonlinear hyperbolic systems of conservation laws. While such singularities are sources of obvious theoretical difficulties—classical solutions should be abandonned to the profit of weak solutions—they are also an obstacle to the convergence of numerical approximation methods, in the sense that they deteriorate the rate of decay of the error with respect to the size of the discrete problem : achieving a prescribed accuracy will typically require finer resolution and therefore heavier computational cost and memory storage, in comparison to the approximation of smooth solutions. Let us remark that singularities

*Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, France. E-mail: cohen@ann.jussieu.fr

often have a physical relevance : they represent the concentration of stress in elasticity, boundary layers in viscous fluid flows, shock waves in gas dynamics... It is therefore a legitimous requirement that they should be accurately resolved by the numerical method.

In this context, the use of adaptive methods, appears as a natural solution to improve the approximation at a reasonable computational cost. Here, the word *adaptivity* has a twofold meaning : (i) the discretization is allowed to be refined only locally, in particular near the singularities of the solution, and (ii) the resolution algorithm uses information gained during a given stage of the computation in order to derive a new refined discretization for the next stage. The most typical example is *adaptive mesh refinement* based on *a-posteriori* error estimates in the finite element context. While these methods have proved to be computationally successful, the theory describing their advantages over their non-adaptive counterpart is far from being complete. In particular, the *rate of convergence* of the adaptive algorithm, which describes the trade-off between the accuracy and complexity of the approximation, is not clearly understood.

In recent years, wavelet bases have been proposed as an alternative to adaptive mesh refinement, motivated by their good performances in data (more specifically image) compression. In wavelet-based adaptive schemes, the set of basis functions which describe the approximate solution is updated at each stage of the computation. Intuitively, the selection of the appropriate basis functions plays a similar role as the selection of the mesh points in adaptive finite element methods, and one could therefore expect similar performances from both approaches. On a more rigorous level, a specific feature of the wavelet approach is the emergence of a sound theoretical setting which allows to tackle fundational questions such as the rate of convergence of the adaptive method.

The goal of this paper is to give some elements of comparison between adaptive wavelet and mesh refinement methods from this perspective. We shall first describe in §2 a general setting which leads us in §3 to a first comparison between wavelets and adaptive finite elements from the point of view of approximation theory. We discuss in §4 the relation between these results and adaptive algorithms for PDE's. After recalling in §5 the classical approach in the finite element context, we present in §6 an adaptive wavelet strategy which has been applied to various problems, and discuss its fundational specificities. Finally, we shall conclude in §7 by pointing out some intrinsic shortcoming of wavelet-based adaptive methods.

2. A general framework

Approximation theory is the branch of mathematics which studies the process of approximating general functions by simple functions such as polynomials, finite elements or Fourier series. It plays therefore a central role in the accuracy analysis of numerical methods. Numerous problems of approximation theory have in common the following general setting : we are given a family of subspaces $(S_N)_{N\geq 0}$ of a

normed space X, and for $f \in X$, we consider the *best approximation error*

$$\sigma_N(f) := \inf_{g \in S_N} \|f - g\|_X. \tag{1}$$

Typically, N represents the number of parameters which are needed to describe an element in S_N, and in most cases of interest, $\sigma_N(f)$ goes to zero as this number tends to infinity. If in addition $\sigma_N(f) \leq CN^{-s}$ for some $s > 0$, we say that f is approximated at rate s.

Given such a setting, the central problem of approximation theory is to *characterize* by some analytic (typically smoothness) condition those functions f which are approximated at some prescribed rate $s > 0$. Another important problem is how to design simple approximation procedures $f \mapsto f_N \in \Sigma_N$ which avoid solving the minimization problem (1), while remaining near optimal in the sense that

$$\|f - f_N\|_X \leq C\sigma_N(f), \tag{2}$$

for some constant C independent of N and f.

As an example, consider approximation by finite element spaces V_h defined from regular conforming partitions $\mathcal{T}_h$ of a domain $\Omega \subset \mathbb{R}^d$ into simplices with uniform mesh size h. The approximation theory for such spaces is quite classical, see e.g. [12], and can be summarized in the following way. If $W^{t,p}$ denotes the classical Sobolev space, consisting of those functions $f \in L^p$ such that $D^\alpha f \in L^p$ for $|\alpha| \leq t$, we typically have

$$f \in W^{t+r,p} \Rightarrow \inf_{g \in V_h} \|f - g\|_{W^{t,p}} \leq Ch^r \tag{3}$$

provided that V_h is contained in $W^{t,p}$ and that V_h has approximation order larger than $t + r$, i.e. contains all polynomials of degree strictly less than $t + r$. Such classical results also hold for fractional smoothness. We can express them in terms of the number of parameters, remarking that $N := \dim(V_h) \sim h^{-d}$, so that if we set $X = W^{t,p}$ and $S_N := V_h$ with $h := N^{-1/d}$, we have obtained

$$f \in W^{t+r,p} \Rightarrow \sigma_N(f) \leq CN^{-r/d}. \tag{4}$$

We have thus identified an analytic condition which ensures the rate $s = r/d$. Note that this is not a characterization (we only have an implication), yet a deeper analysis shows that an "if and only if" result holds if we slightly modify the notion of Sobolev smoothness (using Besov classes, see [13]). In summary, the rate of approximation in $W^{s,p}$ is governed by the approximation order of the V_h spaces, the dimension d and the level of smoothness of f measured in L^p. Let us finally remark that near-optimal approximation procedures can be obtained if we can find a sequence of finite element projectors $P_N : X \mapsto S_N$ such that $\|P_N\|_{X \to X} \leq K$ with K independent of N : in this case, we simply take $f_N = P_N f$ and remark that $\|f - f_N\|_X \leq (1 + K)\sigma_N(f)$.

In the following we shall address the same questions in the cases of adaptive finite element and wavelet approximation. As we shall see, a specific feature to such cases is that the spaces Σ_N are not linear vector spaces.

3. Adaptive finite elements and wavelets

In the adaptive finite element setting, the number of parameters N is proportional to the number of triangles, but for a given budget N the partition $\mathcal{T}$ and the finite element space $V_{\mathcal{T}}$ are allowed to be locally refined in a way which depends on the function f to be approximated. It is therefore natural to define the approximation spaces S_N as

$$S_N := \cup_{\#(\mathcal{T})\leq N} V_{\mathcal{T}}. \tag{5}$$

It should be well understood that the S_N are not linear vector spaces (the sum of two elements does not in general fall in S_N when their triangulation do not match) but any $g \in S_N$ is still described by $\mathcal{O}(N)$ parameters, which encode both its triangulation $\mathcal{T}$ and its coordinates in $V_{\mathcal{T}}$. The requirement of adaptivity has thus led us to the concept of *nonlinear approximation.*

Wavelet bases offer another track toward nonlinear adaptive approximation. The simplest prototype of a wavelet basis is the *Haar system.* Let us describe this system in the case of expanding a function f defined on $[0,1]$: the first component in this expansion is simply the average of f, i.e. the orthogonal projection $\langle f, e_0\rangle e_0$ onto the function $e_0 = \chi_{[0,1]}$. The approximation is then refined into the average of f on the two half intervals of equal size. This refinement amounts in adding the orthogonal projection $\langle f, e_1\rangle e_1$ onto the function $e_1 = \chi_{[0,1/2]} - \chi_{[1/2,1]}$. Iterating this refinement process, we see that the next components have the same form as e_1 up to a change of scale : at refinement level j, we are adding the projection onto the functions

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad k = 0, \cdots, 2^j - 1, \tag{6}$$

where $\psi = e_1$. Since all these functions are orthogonal to the previous ones, letting j go to $+\infty$, we obtain the expansion of f into an orthonormal basis of $L^2([0,1])$

$$f = \sum_{\lambda\in\nabla} f_\lambda \psi_\lambda, \tag{7}$$

with $f_\lambda := \langle f, \psi_\lambda\rangle$. In the above notation λ concatenates the scale and space parameters j and k, and ∇ is the set of all indices (including also the first function e_0). In order to keep track of the scale j corresponding to an index $\lambda = (j,k)$ we shall use the notation $|\lambda| = j$. More general wavelet systems in one or several space dimension are built from similar nested approximation processes, involving e.g. spline functions or finite elements in place of piecewise constant functions (see [23] or [13] for a general presentation).

This brief description suggests that a natural construction of adaptive wavelet approximations is obtained by using only a limited set of indices λ as the scale $|\lambda|$ grows, which depends on the function to be approximated and typically corresponds to those wavelets whose supports are close to its singularities. It is therefore natural to define the approximation spaces S_N as the set of all N terms combinations

$$S_N := \{\sum_{\lambda\in\Lambda} d_\lambda \psi_\lambda \ ; \ \#(\Lambda) \leq N\}. \tag{8}$$

Again this is obviously not a linear space, since we allow to approximate a function by choosing the best N terms which differ from one function to another. Note that we still do have $S_N + S_N = S_{2N}$.

Both adaptive finite element and wavelet framework have obvious similiarities. However, the answer to the two basic questions raised in the previous section—what are the properties of f which govern the decay of $\sigma_N(f)$ and how to compute in a simple way a near optimal approximation of f in Σ_N —is only fully understood in the wavelet framework. Concerning the first question, a striking result by DeVore and his collaborators [24] is the following : with $X := W^{t,p}$, best N-term wavelet approximation satisfies

$$f \in W^{t+r,q} \Rightarrow \sigma_N(f) \leq CN^{-r/d}, \tag{9}$$

with q and r connected by the relation $1/q = 1/p + r/d$, assuming that the multiresolution approximation spaces associated to the wavelet basis are in $W^{t,p}$ and contain the polynomials of degree strictly less than $t + r$.

Such an estimate should be compared with the linear estimate (4) : the same convergence rate is governed by a much weaker smoothness assumption on f since $q < p$ (as in the linear case, an "iff and only if" result can be obtained up to slight technical modifications in the statement of (9)). This result gives a precise mathematical meaning to the spatial adaptation properties of best N-term wavelet approximation: a function f having isolated discontinuity, has usually a smaller amount of smoothness $s + t$ when measured in L^p than when measured in L^q with $1/q = 1/p + t/d$, and therefore $\sigma_N(f)$ might decrease significantly faster than $\varepsilon_N(f)$.

The answer to the second question is given by a result due to Temlyakov : if $f = \sum_{\lambda \in \nabla} d_\lambda \psi_\lambda$, and if we measure the approximation error in $X = W^{t,p}$, a near optimal strategy when $1 < p < \infty$ consists in the thresholding procedure which retains the N largest contributions $\|d_\lambda \psi_\lambda\|_X$: if Λ_N is the corresponding set of indices, one can prove that there exists $C > 0$ independent of N and f such that

$$\|f - \sum_{\lambda \in \Lambda_N} d_\lambda \psi_\lambda\|_X \leq C\sigma_N(f). \tag{10}$$

This fact is obvious when $X = L^2$ using the orthonormal basis property. It is a remarkable property of wavelet bases that it also holds for more general function spaces. In summary, thresholding plays for best N-term wavelet approximation an analogous role as projection for linear finite element approximation.

In the adaptive finite element framework, a similar theory is far from being complete. Partial answers to the basic questions are available if one chooses to consider adaptive partitions with *shape constraints* in terms of a uniform bound on the aspect ratio of the elements

$$\max_{K \in \mathcal{T}} \Big([\mathrm{Diam}(K)]^d / \mathrm{vol}(K) \Big) \leq C. \tag{11}$$

Such a restriction means that the local refinement is isotropic, in a similar way to wavelets. In such a case, we therefore expect a rate of approximation similar to (9). Such a result is not available, yet the following can be proved [13] for Lagrange

finite elements of degree m : if for any given tolerance $\varepsilon > 0$, one is able to build a partition $\mathcal{T} = \mathcal{T}(\varepsilon)$ of cardinality $N = N(\varepsilon)$ such that on each $K \in \mathcal{T}$ the local error of approximation by polynomials satisfies

$$\frac{\varepsilon}{2} \leq \inf_{p \in \Pi_m} \|f - p\|_{W^{t,p}(K)} \leq \varepsilon, \tag{12}$$

then we can build global approximants $f_N \in V_{\mathcal{T}} \subset S_N$ such that

$$f \in W^{t+r,q} \Rightarrow \|f - f_N\|_X \leq CN^{-r/d}, \tag{13}$$

with q and r connected by the relation $1/q = 1/p + r/d$ and assuming $s + t < m$. The effective construction of $\mathcal{T}(\varepsilon)$ is not always feasible, in particular due to the conformity constraints on the partition which does not allow to connect very coarse and very fine elements without intermediate grading. However, this result shows that from an intuitive point of view, the adaptive finite element counterpart to wavelet thresholding amounts in *equilibrating the local error* over the partition. One can actually use these ideas in order to obtain the estimate (9) for adaptive finite elements under the more restrictive assumption that $1/q < 1/p + r/d$. Let us finally mention that the approximation theory for adaptive finite elements without shape constraints is an open problem.

4. Nonlinear approximation and PDE's

Nonlinear approximation theory has opened new lines of research on the theory of PDE's and their numerical discretization. On the one hand, it is worth revisiting the regularity theory of certain PDE's for which the solutions develop singularities but might possess significantly higher smoothness in the scale of function spaces which govern the rate of nonlinear approximation in a given norm than in the scale which govern the rate of linear approximation in the same norm. Results of this type have been proved in particular for elliptic problems on nonsmooth domains [20] and for scalar 1D conservation laws [25]. These results show that if u is the solution of such equations, the rate of decay of $\sigma_N(u)$ is significantly higher for best N-term approximation than for the projection on uniform finite element spaces, therefore advocating for the use of adaptive discretizations of such PDE's.

On the other hand these results also provide with an ideal *benchmark* for adaptive discretizations of the equation, since $\sigma_N(u)$ represents the best accuracy which can be achieved by N parameters. In the wavelet case these parameters are typically the N largest coefficients of the exact solution u. However, in the practice of solving a PDE, these coefficients are not known, and neither is the set Λ corresponding to the indices of the N largest contributions $\|d_\lambda \psi_\lambda\|$. It is therefore needed to develop appropriate *adaptive resolution strategies* as a substitute to the thresholding procedure. Such strategies aim at detecting the indices of the largest coefficients of the solutions and to compute them accurately, in a similar way that adaptive mesh refinement strategies aim at contructing the optimal mesh for finite element approximation. In both contexts, we could hope for an algorithm which builds approximations $u_N \in \Sigma_N$ such that $\|u - u_N\|_X$ is bounded up to a fixed

multiplicative constant by $\sigma_N(u)$ for a given norm of interest, but this requirement is so far out of reach. A more reasonable goal is that the adaptive strategy exhibits the optimal rate of approximation : if $\sigma_N(u) \leq CN^{-s}$ for some $s > 0$, then $\|u-u_N\|_X \leq CN^{-s}$ up to a change in the constant. Another requirement is that the adaptive algorithm should be *scalable*, i.e. the number of elementary operations in order to compute u_N remains proportional to N. Let us finally remark that the norm $\|\cdot\|_X$ for which error estimates can be obtained is often dictated by the nature of the equation (for example $X = H^1$ in the case of a second order elliptic problem) and that additional difficulties can be expected if one searches for estimates in a different norm.

5. The classical approach

The *classical* approach to numerically solving linear and nonlinear partial differential or integral equations $\mathcal{F}(u) = 0$ by the finite element method is typically concerned with the following issues :

(c1) Well-posedness of the equation, i.e. existence, uniqueness and stability of the solution.
(c2) Discretization into a finite element problem $\mathcal{F}_\mathcal{T}(u_\mathcal{T}) = 0$ by the Galerkin method with $u_\mathcal{T} \in V_\mathcal{T}$, analysis of well-posedness and of the approximation error $\|u - u_\mathcal{T}\|_X$.
(c3) Numerical resolution of the finite dimensional system.
(c4) Mesh refinement based on a-posteriori error estimators in the case of adaptive finite element methods.

Several difficulties are associated to each of these steps. First of all, note that the well-posedness of the finite element problem is in general *not* a consequence of the well-posedness of the continuous problem. Typical examples even in the linear case are *saddle point problems.* For such problems, it is well known that, for Galerkin discretizations to be stable, the finite element spaces for the different solution components have to satisfy certain compatibility conditions (LBB or Inf-Sup condition), which are also crucial in the derivation of optimal error estimates. Thus the discrete problem does not necessarily inherit the "nice properties" of the original infinite dimensional problem. Concerning the numerical resolution of the discrete system, a typical source of trouble is its possible *ill-conditioning*, which interferes with the typical need to resort on iterative solvers in high dimension. An additional difficulty occuring in the case of integral equations is the manipulation of matrices which are densely populated.

Finally, let us elaborate more on the adaptivity step. Since more than two decades, the understanding and practical realization of adaptive refinement schemes in a finite element context has been documented in numerous publications [1, 2, 3, 27, 33]. Key ingredients in most adaptive algorithms are *a-posteriori error estimators* which are typically derived from the current residual $\mathcal{F}(u_\mathcal{T})$: in the case where the Frechet derivative $D\mathcal{F}(u)$ is an isomorphism between Banach function spaces X to Y, one can hope to estimate the error $\|u - u_\mathcal{T}\|_X$ by the evaluation of

$\|\mathcal{F}(u_{\mathcal{T}})\|_Y$. The rule of thumb is then to decompose $\|\mathcal{F}(u_{\mathcal{T}})\|_Y$ into computable local error indicators η_K which aim to describe as accurately as possible the local error on each element $K \in \mathcal{T}$. In the case of elliptic problems, these indicators typically consist of local residuals and other quantities such as jumps of derivatives across the interface between adjacent elements. A typical refinement algorithm will subdivide those elements K for which the error indicator η_K is larger than a prescribed tolerance ε resulting in a new mesh $\tilde{\mathcal{T}}$. Note that this strategy is theoretically in accordance with our remarks in §3 on adaptive finite element approximation, since it tends to *equilibrate* the local error. Two other frequently used strategies consist in refining a fixed proportion of the elements corresponding to the largest η_K, or the smallest number of elements K for which the η_K contribute to the global error up to a fixed proportion. It is therefore hoped that the iteration of this process from an initial mesh $\mathcal{T}_0$ will produce optimal meshes $(\mathcal{T}_n)_{n\geq 0}$ in the sense that the associated solutions $u_n := u_{\mathcal{T}_n} \in V_{\mathcal{T}_n}$ converge to u at the optimal rate :

$$\sigma_N(u) \leq CN^{-r} \Rightarrow \|u - u_n\|_X \leq C[\#(\mathcal{T}_n)]^{-r}, \tag{14}$$

up to a change in the constant C. Unfortunately, severe obstructions appear when trying to prove (14) even in the simplest model situations. One of them is that η_K is in general not an estimate by above of the local error, reducing the chances to derive the optimal rate. For most adaptive refinement algorithms, the theoretical situation is actually even worse in the sense that it cannot even be proved that the refinement step actually results in a reduction of the error by a fixed amount and that u_n converges to u as n grows. Only recently [26, 30] have proof of convergence appeared for certain type of adaptive finite element methods, yet without convergence rate and therefore no guaranteed advantage over their non-adaptive counterparts.

6. A new paradigm

Wavelet methods vary from finite element method in that they can be viewed as solving systems that are finite sections of one fixed infinite dimensional system corresponding to the discretization of the equation in the full basis. This observation has led to a *new paradigm* which has been explored in [15] for *linear variational problems.* It aims at closely intertwining the analysis—discretization—solution process. The basic steps there read as follows :

(n1) Well-posedness of the variational problem.
(n2) Discretization into an *equivalent* infinite dimensional problem which is well posed in ℓ^2.
(n3) Devise an iterative scheme for the ℓ_2-problem that exhibits a fixed error reduction per iteration step.
(n4) Numerical realization of the iterative scheme by means of an *adaptive application* of the involved infinite dimensional operators within some dynamically updated accuracy tolerances.

Thus the starting point (n1) is the same. The main difference is that one aims at staying as long as possible with the infinite dimensional problem. Only at the very

end, when it comes to applying the operators in the ideal iteration scheme (n4), one enters the finite dimensional realm. However, the finite number of degrees of freedom is determined at each stage by the adaptive application of the operator, so that at no stage any specific trial space is fixed.

The simplest example is provided by the Poisson equation $-\Delta u = f$ on a domain Ω with homogeneous boundary conditions, for which the variational formulation in $X = H_0^1$ reads : find $u \in X$ such that

$$a(u,v) = L(v), \quad \text{for all } \ v \in X, \tag{15}$$

with $a(u,v) := \int_\Omega \nabla u \nabla v$ and $L(v) := \int_\Omega fv$. The well-posedness for a data $f \in X' = H^{-1}$ is ensured by the Lax-Milgram lemma. In the analysis of the wavelet discretization of this problem, we shall invoke the fact that wavelet bases provide norm equivalence for Sobolev spaces in terms of weighted ℓ^2 norms of the coefficients : if $u = \sum_\lambda u_\lambda \psi_\lambda$, one has

$$\|u\|_{H^s}^2 \sim \sum_\lambda \|u_\lambda \psi_\lambda\|_{H^s}^2 \sim \sum_\lambda 2^{2s|\lambda|} |u_\lambda|^2. \tag{16}$$

We refer to [13] and [21] for the general mechanism allowing to derive these equivalences, in particular for Sobolev spaces on domains with boundary conditions such as H_0^1. Therefore, if we renormalize our system in such a way that $\|\psi_\lambda\|_X = 1$, we obtain the norm equivalence

$$\|u\|_X^2 \sim \|U\|^2, \tag{17}$$

where $U := (u_\lambda)_{\lambda \in \nabla}$ and $\|\cdot\|$ denotes the ℓ^2 norm. By duality, one also easily obtains

$$\|f\|_{X'}^2 \sim \|F\|^2, \tag{18}$$

with $F := (\langle f, \psi_\lambda \rangle)_{\lambda \in \nabla}$. The equivalent ℓ^2 system is thus given by

$$AU = F, \tag{19}$$

where $A(\lambda,\mu) = a(\psi_\lambda, \psi_\mu)$ is a symmetric positive definite matrix which is continuous and coercive in ℓ^2. In this case, a converging infinite dimensional algorithm can simply be obtained by the Richardson iteration

$$U^n := U^{n-1} + \tau(F - AU^{n-1}) \tag{20}$$

with $0 < \tau < 2[\lambda_{\max}(A)]^{-1}$ and $U^0 = 0$, which guarantees the reduction rate $\|U - U^n\| \leq \rho \|U - U^{n-1}\|$ with $\rho = \max\{1 - \tau\lambda_{\min}(A), \tau\lambda_{\max}(A) - 1\}$. Note that renormalizing the wavelet system plays the role of a multiscale preconditioning, similar to multigrid yet operated at the infinite dimensional level.

At this stage, one enters finite dimensional adaptive computation by modifying the Richardson iteration up to a prescribed tolerance according to

$$U^n := U^{n-1} + \tau(\mathbf{COARSE}(F,\varepsilon) - \mathbf{APPROX}(AU^{n-1}, \varepsilon)) \tag{21}$$

where $\|F - \mathbf{COARSE}(F,\varepsilon)\| \leq \varepsilon$ and $\|AU - \mathbf{APPROX}(AU^{n-1},\varepsilon)\| \leq \varepsilon$, and the U^n are now finite dimensional vector supported by adaptive sets of indices Λ_n. The

procedure **COARSE**, which simply corresponds to thresholding the data vector F at a level corresponding to accuracy ε, can be practically achieved without the full knowledge of F by using some a-priori bounds on the size of the coefficients $\langle f, \psi_\lambda \rangle$, exploiting the local smoothness of f and the oscillation properties of the wavelets. The procedure **APPROX** deserves more attention. In order to limitate the spreading effect of the matrix A, one invokes its *compressibility* properties, namely the possibily to truncate it into a matrix A_N with N non-zero entries per rows and columns in such a way that

$$\|A - A_N\|_{\ell^2 \to \ell^2} \leq CN^{-s}. \tag{22}$$

The rate of compressibility s depends on the available a-priori estimates on the off-diagonal entries $A(\lambda, \mu) := \int_\Omega \nabla \psi_\lambda \nabla \psi_\mu$ which are consequences of the smoothness and vanishing moment properties of the wavelet system, see [14]. Once these properties are established, a first possibility is thus to choose

$$\mathbf{APPROX}(AU^{n-1}, \varepsilon) = A_N U^{n-1} \tag{23}$$

with N large enough so that accuracy ε is ensured. Clearly the modified iteration (21) satisfies $\|U - U^n\| \leq \rho \|U - U^{n-1}\| + 2\tau\varepsilon$, and therefore ensures a fixed reduction rate until the error is of the order $\frac{2\tau}{1-\rho}\varepsilon$, or until the residual $F - AU^n$ is of order $\frac{2\tau\|A\|}{1-\rho}\varepsilon$. A natural idea is therefore to update dynamically the tolerance ε, which is first set to 1 and divided by 2 each time the approximate residual $\mathbf{COARSE}(F, \varepsilon) - \mathbf{APPROX}(AU^{n-1}, \varepsilon)$ is below $[\frac{2\tau\|A\|}{1-\rho} + 3]\varepsilon$ (which is ensured to happen after a fixed number of steps).

We therefore obtain a converging adaptive strategy, so far without information about the convergence rate. It turns out that the optimal convergence rate can also be proved, with a more careful tuning of the adaptive algorithm. Two additional ingredients are involved in this tuning.

Firstly, the adaptive matrix vector multiplication **APPROX** has to be designed in a more elaborate way than (23) which could have the effect of inflating too much the sets Λ_n. Instead, one defines for a finite length vector V

$$\mathbf{APPROX}(AV, \varepsilon) = \sum_{l=0}^{j} A_{2^{j-l}}[V_{2^l} - V_{2^{l-1}}] \tag{24}$$

where V_N denotes the restriction of V to its N largest components (with the notation $V_{1/2} = 0$), and j is the smallest positive integer such that the residual $\sum_{l=0}^{j} \|A - A_{2^{j-l}}\| \|V_{2^l} - V_{2^{l-1}}\| + \|A\| \|V - V_{2^j}\|$ is less than ε. In this procedure, the spreading of the operator is more important on the largest coefficients which are less in number, resulting in a significant gain in the complexity of the outcome.

Secondly, additional coarsening steps are needed in order to further limitate the spreading of the sets Λ_n and preserve the optimal rate of convergence. More precisely, the procedure **COARSE** is applied to U^n with a tolerance proportional to ε, for those n such that ε will be updated at the next iteration.

With such additional ingredients, it was proved in [15] that the error has the optimal rate of decay in the sense that

$$\sigma_N(u) \leq CN^{-s} \Rightarrow \|u - u_n\|_X \sim \|U - U^n\| \leq C[\#(\Lambda_n)]^{-s}, \tag{25}$$

and that moreover, the computational cost of producing u_n remains proportional to $\#(\Lambda_n)$. It is interesting to note that this strategy extends to non-elliptic problems such as saddle-points problems, without the need for compatibility conditions, since one inherits the well-posedness of the continuous problem which allows to obtain a converging infinite dimensional iteration, such the Uzawa algorithm or a gradient descent applied to the least-square system (see also [15, 19]). The extension to nonlinear variational problems, based on infinite dimensional relaxation or Newton iterations, has also been considered in [16]. It requires a specific procedure for the application of the nonlinear operator in the wavelet coefficients domain which generalizes (23). It should also be mentioned that matrix compressibility also applies in the case of integral operators which have quasi-sparse wavelet discretizations. Therefore several of the obstructions from the classical approach—conditioning, compatibility, dense matrices—have disappeared in the wavelet approach.

Let us finally mention that the coarsening steps are not really needed in the practical implementations of the adaptive wavelet method (for those problems which have been considered so far) which still does exhibit optimal convergence rate. However, we do not know how to prove (25) without these coarsening steps. There seems to be a similar situation in the finite element context : it has recently been proved in [10] that (14) can be achieved by an adaptive mesh refinement algorithm which incorporates coarsening steps, while these steps are not needed in practice.

7. Conclusions and shortcomings

There exist other approaches for the development of efficient wavelet-based adaptive schemes. In particular, an substantial research activity has recently been devoted to *multiresolution adaptive processing* techniques, following the line of idea introduced in [28, 29]. In this approach, one starts from a classical and reliable scheme on a uniform grid (finite element, finite difference or finite volume) and applies a discrete multiresolution decomposition to the numerical data in order to *compress* the computational time and memory space while preserving the accuracy of the initial scheme. Here the adaptive sets Λ_n are therefore limited within the resolution level of the uniform grid where the classical scheme operates. This approach seems more appropriate for hyperbolic initial value problems [17, 22], in which a straightforward wavelet discretization might fail to converge properly. It should again be compared to its adaptive mesh refinement counterpart such as in [6, 5].

Let us conclude by saying that despite its theoretical success, in the sense of achieving for certain classes of problems the optimal convergence rate with respect to the number of degrees of freedom, the wavelet-based approach to adaptive numerical simulation suffers from three major curses.

The curse of geometry : while the construction of wavelet bases on a rectangular domains is fairly simple—one can use tensor product techniques and inherit the simplicity of the univariate construction—it is by far less trivial for domains with complicated geometries. Several approaches have been proposed to deal with this situation, in particular domain decomposition into rectangular patches or hierarchical finite element spaces, see [13, 21], and concrete implementations are nowaday available, but they result in an unavoidable loss of structural simplicity in comparison to the basic Haar system of §3.

The curse of data structure : encoding and manipulating the adaptive wavelet approximations U^n to the solution means that we both store the coefficients and the indices of the adaptive set Λ_n which should be dynamically updated. The same goes for the indices of the matrix A which are used in the matrix-vector algorithm (24) at each step of the algorithm. This dynamical adaptation, which requires appropriate data structure, results in major overheads in the computational cost which are observed in practice : the numerical results in [4] reveal that while the wavelet adaptive algorithm indeed exhibits the optimal rate of convergence and slightly outperforms adaptive finite element algorithms from this perspective, the latter remains significantly more efficient from the point of view of computational time.

The curse of anisotropy : adaptive wavelet approximation has roughly speaking the same properties as isotropic refinement. However, many instances of singularities such as boundary layers and shock waves, have anisotropic features which suggests that the refinement should be more pronounced in one particular direction. From a theoretical point of view, the following example illustrate the weakness of wavelet bases in this situation : if $f = \chi_\Omega$ with $\Omega \subset \mathbb{R}^d$ a smooth domain, then the rate of best N-term approximation in $X = L^2$ is limited to $r = 1/(2d-2)$ and therefore deteriorates as the dimension grows. Wavelet bases should therefore be reconsidered if one wants to obtain better rates which take some advantage of the geometric smoothness of the curves of dicsontinuities. On the adaptive finite element side, anisotropic refinement has been considered and practically implemented, yet without a clean theory available for the design of an optimal mesh.

The significance of wavelets in numerical analysis remains therefore tied to these curses and future breakthrough are to be expected once simple and appropriate solutions are proposed in order to deal with them.

References

[1] Babushka, I. and W. Reinhbolt (1978) *A-posteriori analysis for adaptive finite element computations*, SIAM J. Numer. Anal. 15, 736–754.

[2] Babuška, I. and A. Miller (1987), *A feedback finite element method with a-posteriori error estimation: Part I. The finite element method and some basic properties of the a-posteriori error estimator*, Comput. Methods Appl. Mech.

Engrg. **61**, 1–40.
[3] Bank, R.E. and A. Weiser (1985), *Some a posteriori error estimates for elliptic partial differential equations*, Math. Comp., **44**, 283–301.
[4] Barinka, A., T. Barsch, P. Charton, A. Cohen, S. Dahlke, W. Dahmen, K. Urban (1999), *Adaptive wavelet schemes for elliptic problems—Implementation and numerical experiments, IGPM Report # 173 RWTH Aachen*, SIAM J. Sci. Comp. **23**, 910–939.
[5] Berger, M. and P. Collela (1989) *Local adaptive mesh refinement for shock hydrodynamics*, J. Comp. Phys. 82, 64–84.
[6] Berger, M. and J. Oliger (1984) *Adaptive mesh refinement for hyperbolic partial differential equations*, J. Comp. Phys. 53, 482–512.
[7] Bertoluzza, S. (1995) *A posteriori error estimates for wavelet Galerkin methods*, Appl. Math. Lett. 8, 1–6.
[8] Bertoluzza, S. (1997) *An adaptive collocation method based on interpolating wavelets*, in *Multiscale Wavelet Methods for PDEs*, W. Dahmen, A. J. Kurdila, P. Oswald (eds.), Academic Press, 109–135.
[9] Bihari, B. and A. Harten (1997) *Multiresolution schemes for the numerical solution of 2–D conservation laws*, SIAM J. Sci. Comput. **18**, 315–354.
[10] Binev, P., W. Dahmen and R. DeVore (2002), *Adaptive finite element methods with convergence rates*, preprint IGPM-RWTH Aachenm, to appear in Numerische Mathematik.
[11] Canuto, C. and I. Cravero (1997), *Wavelet-based adaptive methods for advection-diffusion problems*, Math. Mod. Meth. Appl. Sci. **7**, 265–289.
[12] Ciarlet, P.G. (1991), *Basic error estimates for the finite element method*, Handbook of Numerical Analysis, vol II, P. Ciarlet et J.-L. Lions eds., Elsevier, Amsterdam.
[13] Cohen, A. (2000), *Wavelets in numerical analysis*, Handbook of Numerical Analysis, vol. VII, P.G. Ciarlet and J.L. Lions, eds., to appear in 2003 as a book"Numerical analysis of wavelet methods", Elsevier, Amsterdam.
[14] Cohen, A., W. Dahmen and R. DeVore (2000), *Adaptive wavelet methods for elliptic operator equations—convergence rate*, Math. Comp. **70**, 27–75.
[15] Cohen, A., W. Dahmen and R. DeVore (2002), *Adaptive wavelet methods for operator equations—beyond the elliptic case*, Found. of Comp. Math. **2**, 203–245.
[16] Cohen, A., W. Dahmen and R. DeVore (2002), *Adaptive wavelet methods for nonlinear variational problems*, preprint IGPM-RWTH Aachen, submitted to SIAM J. Num. Anal.
[17] Cohen, A., S.M. Kaber, S. Mueller and M. Postel (2002), *Fully adaptive multiresolution finite volume schemes for conservation laws*, Math. of Comp. **72**, 183–225.
[18] Cohen, A. and R. Masson (1999), *Wavelet adaptive methods for elliptic problems—preconditionning and adaptivity*, SIAM J. Sci. Comp. 21, 1006–1026.
[19] Dahlke, S., W. Dahmen and K. Urban (2001), *Adaptive wavelet methods for saddle point problems—optimal convergence rates*, preprint IGPM-Aachen
[20] Dahlke, S. and R. DeVore (1997), *Besov regularity for elliptic boundary value*

problems, Communications in PDEs 22, 1–16.
[21] Dahmen, W. (1997), *Wavelet and multiscale methods for operator equations*, Acta Numerica 6, 55–228.
[22] Dahmen, W., B. Gottschlich-Müller and S. Müller (2001), *Multiresolution Schemes for Conservation Laws*, Numerische Mathematik 88, 399–443.
[23] Daubechies, I. (1992), *Ten lectures on wavelets*, SIAM, Philadelphia.
[24] DeVore, R. (1997), *Nonlinear Approximation*, Acta Numerica 51–150.
[25] DeVore, R. and B. Lucier (1990), *High order regularity for conservation laws*, Indiana J. Math. 39, 413–430.
[26] Dörfler, W. (1996), *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Num. Anal. **33**, 1106–1124
[27] Eriksson, K., D. Estep, P. Hansbo, and C. Johnson (1995), *Introduction to adaptive methods for differential equations*, Acta Numerica **4**, Cambridge University Press, 105–158.
[28] Harten, A. (1994), *Adaptive multiresolution schemes for shock computations*, J. Comp. Phys. 115, 319–338.
[29] Harten, A. (1995), *Multiresolution algorithms for the numerical solution of hyperbolic conservation laws*, Comm. Pure and Appl. Math. 48, 1305–1342.
[30] Morin, P., R. Nocetto and K. Siebert (2000), *Data oscillation and convergence of adaptive FEM*, SIAM J. Num. Anal. **38**, 466–488.
[31] Petrushev, P. (1988), *Direct and converse theorems for spline and rational approximation and Besov spaces*, in *Function spaces and applications*, M. Cwikel, J. Peetre, Y. Sagher and H. Wallin, eds., Lecture Notes in Math. 1302, Springer Verlag, Berlin, 363–377.
[32] Temlyakov, V. (1998), *Best N-term approximation and greedy algorithms*, Adv. Comp. Math. **8**, 249–265.
[33] Verfürth, R. (1994), *A-posteriori error estimation and adaptive mesh refinement techniques*, Jour. Comp. Appl. Math. 50, 67–83.

ICM 2002 · Vol. I · 621–630

Energy Landscapes and Rare Events

Weinan E* Weiqing Ren† Eric Vanden-Eijnden‡

Abstract

Many problems in physics, material sciences, chemistry and biology can be abstractly formulated as a system that navigates over a complex energy landscape of high or infinite dimensions. Well-known examples include phase transitions of condensed matter, conformational changes of biopolymers, and chemical reactions. The energy landscape typically exhibits multiscale features, giving rise to the multiscale nature of the dynamics. This is one of the main challenges that we face in computational science. In this report, we will review the recent work done by scientists from several disciplines on probing such energy landscapes. Of particular interest is the analysis and computation of transition pathways and transition rates between metastable states. We will then present the string method that has proven to be very effective for some truly complex systems in material science and chemistry.

2000 Mathematics Subject Classification: 60-08, 60F10, 65C.
Keywords and Phrases: Energy landscapes, Stochastic effects, Rare events, Transition pathways, Transition rates, String method.

1. Introduction

Many problems in biology, chemistry and material science can be formulated as the study of the energy or free energy landscape of the underlying system. Well-known examples of such problems include the conformational changes of macromolecules, chemical reactions and nucleation in condensed systems. Very often the dimension of the state space is very large, and the energy landscape exhibits a hierarchy of structures and scales. These problems are becoming a major challenge in their respective scientific disciplines and are beginning to receive attention from

*Department of Mathematics and PACM, Princeton University, Fine Hall, Princeton, NJ 08544, USA and School of Mathematics, Peking University, Beijing, 100871, China. E-mail: weinan@math.princeton.edu

†Courant Institute of Mathematical Sciences, New York University, New York 10012, USA. Email: weiqing@cims.nyu.edu

‡Courant Institute of Mathematical Sciences, New York University, New York 10012, USA. Email: eve2@cims.nyu.edu

the mathematics community. In this article, we report recent work in this direction. For a detailed account, we refer to [4, 5, 6, 7, 12].

We begin with a simple example. Plotted in Figure 1 is the solution of the stochastic differential equation

$$dx(t) = -\nabla_x V(x(t))dt + \sqrt{\varepsilon}dW(t) \tag{1.1}$$

where the potential

$$V(x) = \frac{1}{4}(1-x^2)^2 \tag{1.2}$$

and $dW(t)$ is Gaussian white noise, $\varepsilon = 0.06, x(0) = -1$. Without the random perturbation, the solution would be $x(t) \equiv x(0) = -1$. Indeed the deterministic part of dynamics in (1.1) does nothing but taking the system to local equilibrium states. With the random perturbation, the solution, over long time, exhibits completely different behavior. It fluctuates around the two local minima of $V, x = -1$ and 1, with sudden transitions between these two states. The time scale of the transition, t_M is much larger than the time scale of the fluctuation around the local minima, t_R. For this reason, we refer to $x = -1$ and 1 as the metastable states.

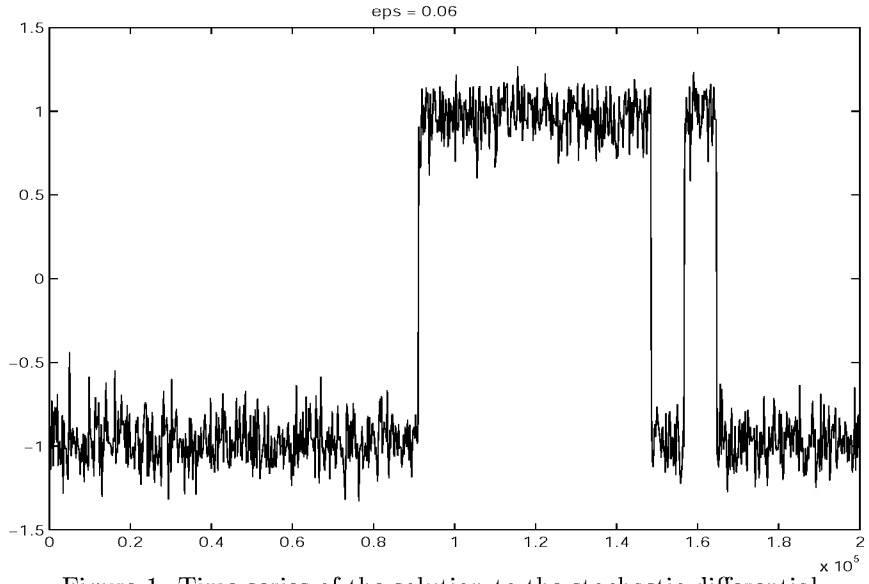


Figure 1. Time series of the solution to the stochastic differential equation (1.1), with $\varepsilon = 0.06$.

Obviously the transition between the metastable states is of more interest than the local fluctuation around them. The transition time is much larger since it requires the system to overcome the energy barrier between the two states. This is only possible because of the noise. When ε is small, a huge noise term is required

to accomplish this. For this reason, such events are very rare, and this is the origin of the disparity between the time scales t_M and t_R.

This simple example illustrates one of the major difficulties in modeling such systems, namely the disparity of the time scales. It does not, however, illustrate the other major difficulty, namely, the large dimension of the state space and the complexity of the energy landscapes. Indeed for typical systems of interests the energy landscape can be very complex. There can be a huge number of local minima in the state space. The usual concept of hopping over barriers via saddle points may not apply (see [3]).

In applications, the noise comes typically from thermal noise. In this case, we should note that even though the potential energy landscapes might be rough and contain small scale features, the system itself experiences a much smoother landscape, the free energy landscape, since some of the small scale features on the potential energy landscape are smoothed out by the thermal noise.

Our objective in modeling such systems are the following:

1. Find the transition mechanism between the metastable states.
2. Find the transition rates.
3. Reduce the original dynamics to the dynamics of a Markov chain on the metastable states.

Our discussion will be centered around the following model problems:

$$\gamma \dot{x}(t) = -\nabla V(x(t)) + \sqrt{\varepsilon}\dot{W}(t) \tag{1.3}$$

or

$$m\ddot{x}(t) + \gamma \dot{x}(t) = -\nabla V(x(t)) + \sqrt{m\varepsilon}\dot{W}(t) \tag{1.4}$$

ε is related to the temperature of the system by $\varepsilon = 2\gamma k_B T$ where k_B is the Boltzmann constant. We refer to (1.3) as type-I gradient flow and (1.4) as type-II gradient flow.

Before proceeding further, let us remark that there is a very well-developed theory, the large-deviation theory, or the Wentzell-Freidlin theory [8], that deals precisely with questions of the type that we discussed above. However as was explained in [7, 12], this theory is not best suited for numerical purpose. Therefore we will seek an alternative theoretical framework that is more useful for numerical computations.

2. Transition state theory

Transition state theory (TST) [9] has been the classical framework for addressing the questions we are interested in. It assumes the existence and explicit knowledge of a reaction coordinate, denoted by q, that connects the two metastable states. In addition it assumes that along the reaction coordinate there exists a well-defined transition state, which is typically the saddle point configuration, say at $q = 0$, and the two regions $\{q < 0\}$ and $\{q > 0\}$ defines the two metastable regions A and B. For these reasons, transition state theory is restricted to cases when the

system is simple and the energy landscape is smooth, i.e. the energy barriers are larger than the thermal energy k_BT.

Knowing the transition state, TST calculates the transition rates by placing particles at the transition state, and measuring the flux that goes into the two regions. For example, the transition rate from A to B is given approximately by

$$k_{A\to B} = \frac{1}{Z_0}\int \dot{q}(t)\delta_\Gamma(q(t))\theta(q(t))d\mu_A(q(0)) \tag{2.1}$$

where

$$Z_0 = \int d\mu_A(q(0)) \tag{2.2}$$

Here δ_Γ is the surface delta function at $q = 0$, θ is the Heaviside function, μ_A is the Lebesgue measure restricted to A. For a system with a single particle of mass m and potential V, this gives [9]

$$k_{A\to B} = \frac{\omega_0}{2\pi}e^{-\frac{\delta E}{k_BT}} \tag{2.3}$$

where δE is the energy barrier at the transition state, $\omega_0 = \left(\frac{V''(x_A)}{m}\right)^{\frac{1}{2}}$, x_A denotes the location of the local minimum inside A. Formulas such as (2.3) are the origin of the Arrhenius law for chemical reaction rates and Boltzmann factor for hopping rates in kinetic Monte Carlo models.

3. Reduction to Markov chains on graphs

For simplicity, we will discuss mainly type-I gradient flows (1.3). The Fokker-Planck equation can be expressed as

$$\frac{\partial p}{\partial t}(x,t) = \nabla\cdot\left(p_s(x)\nabla\left(\frac{p(x,t)}{p_s(x)}\right)\right) \tag{3.1}$$

where p_s is the equilibrium distribution

$$p_s(x) = \frac{1}{Z}e^{-\frac{V(x)}{k_BT}}$$

Z is the normalization constant $Z = \int_{R^n} e^{-\frac{V(x)}{k_BT}}dx$. The states of the Markov chain consist of the sets $\{B_j\}_{j=1}^J$, where $\{B_j\}_{j=1}^J$ satisfies

1. The B_j's are mutually disjoint
2. $$\int_B p_s(x)dx = 1 + o(k_BT) \tag{3.2}$$

where $B = \cup_{j=1}^J B_j$. An illustration of the collection the $\{B_j\}_{j=1}^J$ is given in Figure 2. $\{B_j\}$ depends on T. As T decreases, the B_j's exhibits a hierarchical structure.

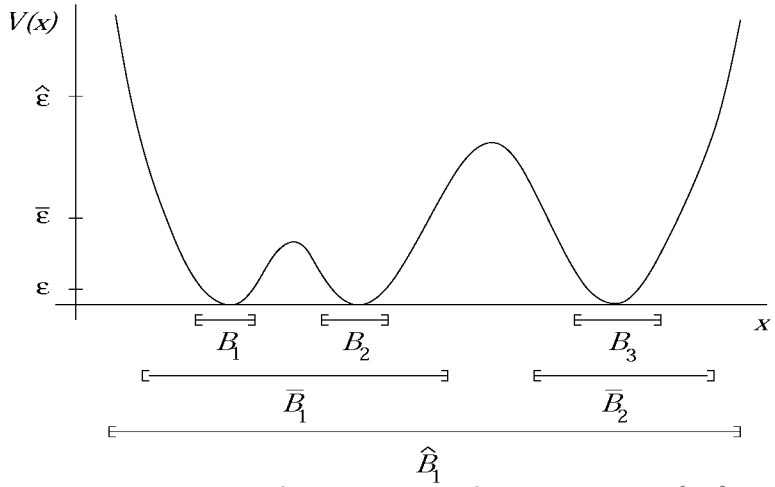


Figure 2. Illustration of the collection of metastable sets $\{B_j\}$ at different energies.

Having defined the states of the Markov chain, we next compute the transition rates between neighboring states. Denote by A and B two such neighboring states. We would like to compute the transition rate from A to B. Without loss of generality, we may assume $J = 2$. Let B_1, B_2 be the metastable region containing A and B respectively, and let $n_j(t) = \int_{B_j} p(x,t)dx$, $N_j = \int_{B_j} p_s(x)dx$, $j = 1, 2$. Applying Laplace's method to (3.1) we get [7]

$$\dot{n}_j(t) = \frac{\varepsilon}{\kappa}\left(\frac{n_2(t)}{N_2} - \frac{n_1(t)}{N_1}\right) + \text{higher order terms} \tag{3.3}$$

where

$$\kappa = \int_0^1 d\alpha \left(\int_{S^0(\alpha)} p_s(x)dx\right)^{-1} |\varphi^0(\alpha)| \tag{3.4}$$

$\{\varphi^0(\alpha), 0 \le \alpha \le 1\}$ is a so-called minimal energy path, to be defined below, $\{S^0(\alpha)\}$ is the family of hyperplanes normal to φ^0.

The minimal energy path (MEP) is defined as follows. If V is smooth, then φ^0 is a MEP if

$$(\nabla V)^{\perp}(\varphi^0(\alpha)) = 0 \tag{3.5}$$

for all $\alpha \in [0,1]$, i.e. ∇V restricted to φ^0 is parallel to φ^0. In general there is not a unique φ^0 that is particularly significant, but rather a collection (a tube or several tubes) of paths contribute to the transition rates. However, one can define a MEP self-consistently via the equation

$$\varphi^0(\alpha) = \frac{1}{Z(\alpha)} \int_{S^0(\alpha)} x e^{-\frac{V(x)}{k_B T}} dx \tag{3.6}$$

where $Z(\alpha) = \int_{S^0(\alpha)} e^{-\frac{V(x)}{k_B T}} dx$. In the case when V has two scales: $V = \bar{V} + \delta V$, $|\delta V| \leq O(k_B T)$, and $\bar{V}$ is smooth, then φ^0 can be defined as the MEP of $\bar{V}$.

For type-I gradient systems, if the two metastable sets are separated by a single saddle point, then the MEP is the unstable manifold associated with the saddle point. MEP for type-II gradient systems is less trivial. In this case (3.3)-(3.6) have to be modified [7]. Consider the simple example

$$\begin{cases} \dot{q} &= p \\ \dot{p} &= -\frac{\partial V}{\partial q}(q) - p + \sqrt{\varepsilon}\dot{W} \end{cases}$$

where $V(q) = \frac{1}{4}(1-q^2)^2$. It has two local equilibrium states $A = (-1, 0)$ and $B = (1, 0)$. The MEP that connects A and B is plotted in Figure 3. It is not a smooth curve. The velocity is reversed at the saddle point. To verify that the MEP does reflect the true behavior of the transition path, we also plot the transition path obtained from direct simulation of the stochastic differential equation.

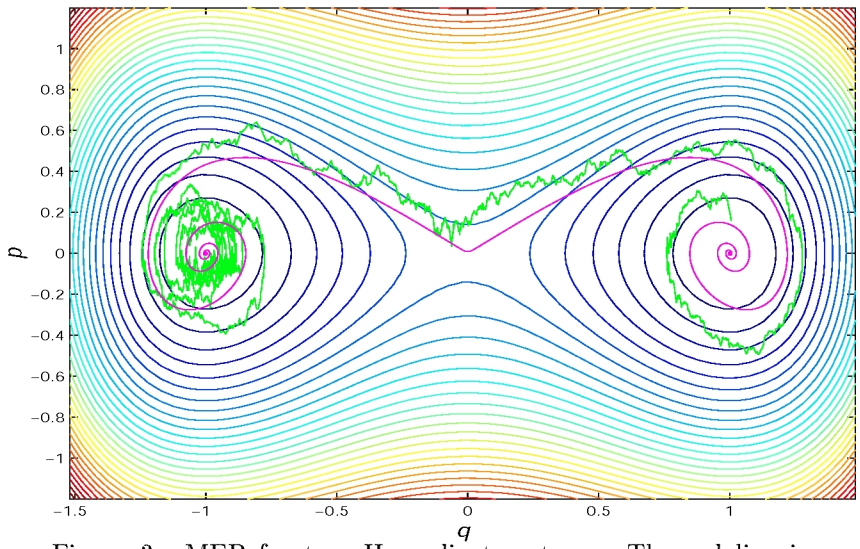


Figure 3. MEP for type-II gradient systems. The red line is the MEP in phase space, the green line is the transition path computed from solving the stochastic differential equation.

MEP is a very important concept since it defines the "most probable" transition path from which transition rates can be computed via equation (3.3). However we should emphasize that from a numerical point of view our task is not that of a conventional optimization or control problem, since there is not an objective function that we can easily work with. Instead our aim is to perform importance sampling in path space to sample the paths that contribute significantly to the switching.

Finally, if there exists a MEP that connects two metastable sets without going through a third one, then we connect these two metastable sets by a link. In this way, we form a graph. The original dynamics is then reduced to a Markov chain on this graph.

4. Previous numerical techniques

A variety of numerical techniques have been developed, most prominently in chemistry, but also in biology and material science, for computing MEPs and sometimes transition rates. Among the most well-known techniques in the chemistry literature are the nudged elastic band method and the transition path sampling technique. The nudged elastic band method (NEB) [10] aims at computing the MEP defined by (3.5). It represents the MEP by a discrete chain of states. These states evolve according to the potential forces of the system. To prevent the states from all falling to the two local equilibrium states, a spring force is applied to neighboring states to penalize the non-uniformity in the distribution of the states along the chain. This by itself may cause convergence to a path which is not a MEP. Hence a nudging technique is used, namely only the normal component of the potential force and the tangential component of the spring force is applied. NEB is a very effective method for small systems with relatively smooth energy landscapes. It has two main drawbacks. One is that it is highly inefficient and may not even be applicable to systems with rough energy landscapes. The other is the choice of the elastic constant. A large elastic constant requires a small time step in the evolution of the states. A small elastic constant will not achieve the desired uniformity of the states and hence will not give the required accuracy for the energy barrier.

A second important technique is transition path sampling (TPS) [2, 3]. This method aims at complex systems with rough energy landscapes by developing a Monte Carlo technique that samples the path space. Its efficiency hinges on the ability to produce new accepted paths from old ones.

Other techniques include the ridge method, blue mooth sampling, etc. [1]. Often these methods require knowing beforehand the reaction coordinate.

Elber et. al propose to minimize the Onsager-Machlup action as a way of finding the most probable path for macromolecular systems [11]. The Onsager-Machlup action is the same as the Wentzell-Freidlin action. From a numerical point of view, there are certain difficulties associated with minimizing this action functional. These issues are discussed in [7].

5. The string method

The basic idea in the string method, developed in [4, 5, 7], is to represent transition paths by their intrinsic parameterization in order to efficiently evolve and sample paths in path space. It has two versions. The zero temperature version is designed for smooth energy landscapes. The finite temperature version is designed

for rough energy landscapes in which case thermal noise acts to smooth out the small scale features.

The simplest example of a zero-temperature string method is to evolve curves in path space by the gradient flow

$$\varphi_t(\alpha,t) = -(\nabla V)^\perp(\varphi(\alpha,t)) + r(\alpha,t)\hat{\tau}(\alpha,t) \tag{5.1}$$

Here $\hat{\tau}$ is the tangent vector of the curve $\{\varphi(\cdot,t)\}$, $(\nabla V)^\perp(\varphi)$ denotes the component of ∇V normal to $\hat{\tau}$, r is the Lagrange multiplier that enforces certain specific parameterization of the curves. For example if we require equal arclength parameterization, then we need $\frac{\partial}{\partial\alpha}|\varphi_\alpha| = 0$, i.e.

$$r(\alpha,t) = \alpha\int_0^1 \nabla V(\varphi(\alpha',t))\cdot\hat{\tau}(\alpha',t)d\alpha' - \int_0^\alpha \nabla V(\varphi(\alpha',t))\cdot\hat{\tau}(\alpha',t)d\alpha' \tag{5.2}$$

We call such curves with intrinsic parameterization strings.

In practice the strings are discretized into a collection of points. These points move according to the normal component of the potential force. A reparameterization step is applied once in a while to enforce the proper parameterization of the strings.

The finite temperature string method is designed for systems with rough energy landscapes, particularly the case when the potential can be expressed in the form

$$V(x) = \bar{V}(x) + \delta V(x)$$

where $\bar{V}$ is smooth and $|\delta V| \le O(k_BT)$. In this case we would like to compute the MEP of $\bar{V}$ without first computing $\bar{V}$ explicitly. This is achieved by creating an ensemble of a special type. Our computational object will be a string connecting the two metastable sets, together with a family of probability measures on the hyperplanes normal to the string. Consider the stochastic equation

$$\varphi_t^\omega(\alpha,t) = -\mathbf{P}_\alpha^0(\nabla V(\varphi^\omega(\alpha,t))) + r^0(\alpha,t)\hat{\tau}^0(\alpha,t) + \mathbf{P}_\alpha^0\eta^\omega(\alpha,t) \tag{5.3}$$

where η^ω is Gaussian noise with mean 0 and correlation

$$\mathbf{E}\eta^\omega(\alpha,t)\eta^\omega(\alpha',\tau) = \begin{cases} 2k_BT\delta(t-\tau), & \text{if } \alpha = \alpha' \\ 0, & \text{if } \alpha \ne \alpha' \end{cases}$$

The projection operator $\mathbf{P}_\alpha^0$ is defined by projecting to the hyperplane normal to the string $\{\varphi^0(\alpha,t), 0 \le \alpha \le 1\}$, where

$$\varphi^0(\alpha,t) = \mathbf{E}\varphi^\omega(\alpha,t)$$

$\hat{\tau}^0(\alpha,t)$ is the tangent vector of φ^0 at α, r^0 is the Lagrange multiplier that enforces proper parameterization of φ^0.

Theorem 1.1.

1. *The statistical steady state of* (5.3) *satisfies:*

$$\varphi^0(\alpha) = \frac{1}{Z(\alpha)}\int_{S^0(\alpha)} xe^{-\frac{V(x)}{k_BT}}dx \tag{5.4}$$

where $S^0(\alpha)$ is the hyperplane normal to φ^0 at α, $Z(\alpha) = \int_{S^0(\alpha)} e^{-\frac{V(x)}{k_BT}}dx$.

2. The stationary distribution of (5.3) *is given by the family of distributions*

$$\mu_\alpha(x) = \frac{1}{Z(\alpha)} e^{-\frac{V(x)}{k_B T}} \delta_{S^0(\alpha)}(x) \tag{5.5}$$

Knowing $\{\varphi^0\}$ and $\{\mu_\alpha\}$, the transition rates and free energy landscapes can be computed, see [7].

The finite temperature string method is applied to the perturbed Mueller potential

$$V(x) = V_m(x) + \delta V(x)$$

where $V_m(x)$ is the so-called Mueller potential (see [11]), δV is a random perturbation. The results of the (finite temperature) string method is shown in Figure 4. Also plotted is the MEP of V_m (since $\bar{V} = V_m$ is explicitly known for this particular example) as well as the fluctuations around φ^0.

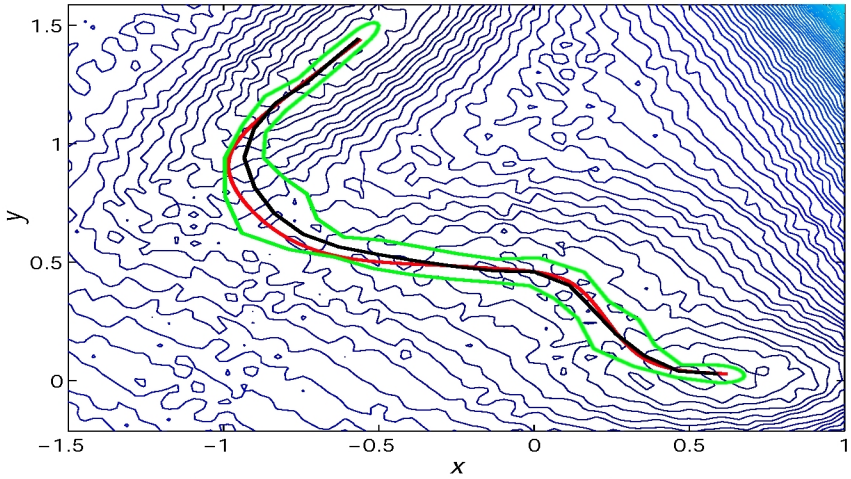


Figure 4. Effective MEP and local fluctuations for the perturbed mueller potential. The red curve is the MEP for $\bar{V} = V_m$. The black curve is the MEP computed from the finite temperature string method. The green curves show the size of the fluctuations.

6. Concluding remarks

There are several important topics that we did not cover in this brief report. These include the effect of dynamics, non-gradient systems, and acceleration techniques. These are discussed in [4, 5, 7, 12]. Also found in these references are applications of the ideas discussed here to thermal activated reversal of magnetic thin films, models of martensitic transformations, and the formation of C_{60} from 60 carbon atoms. The last example is a case when the barrier is entropic. Even though the potential energy is mainly going downhill, the free energy has barriers because of entropic effects. Such examples are found frequently in biopolymers.

From a numerical point of view, our main idea for overcoming the difficulty caused by the disparity of the times scales is to reformulate the problem as a boundary value problem instead of initial value problem, since we have some knowledge of the initial and final state of the system. Compared with other existing methods that assume explicit knowledge of a reaction coordinate, our method finds the reaction coordinate self-consistently during the computation.

The topic discussed here is relatively new in applied mathematics, but it is of paramount importance in science and particularly computational science. Progress in this area will likely have a fundamental impact in many areas of applications.

Acknowledgment. We are grateful to Bob Kohn for many stimulating discussions. This work is partially supported by NSF via grant DMS01-30107.

References

[1] Carter, E. A.; Cicotti, G.; Hynes, J. T.; Kapral, R., Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.* **156** (1989), 472–477.

[2] Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108** (1998), 1964–1977.

[3] Dellago, C.; Bolhuis, P. G.; Geissler, P. L. Transition Path Sampling, Submitted to: *Adv. Chem. Phys.* (2001).

[4] E, W.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B*, in press (2001).

[5] E, W.; Ren, W.; Vanden-Eijnden, E. Energy Landscape and Thermally Activated Switching of Submicron-sized Ferromagnetic Elements. *J. App. Phys.*, submitted.

[6] E, W.; Ren, W.; Vanden-Eijnden, E. Probing Multi-Scale Energy Landscapes Using the String Method. *Phys. Rev. Lett.*, submitted.

[7] E, W.; Ren, W.; Vanden-Eijnden, E. Transition pathways in complex systems. Theory and numerical methods. In preparation.

[8] Freidlin, M. I.; Wentzell, A. D. *Random Perturbations of Dynamical Systems*, 2nd ed. Springer, 1998.

[9] Hänggi, P.; Talkner, P.; Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Physics* **62** (1990), 251–341.

[10] Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In: Classical and Quantum Dynamics in Condensed Phase Simulations. Edited by: Berne, B. J.; Cicotti, G.; Coker, D. F. World Scientific, 1998.

[11] Olender, R.; Elber, R. Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *J. Chem. Phys.* **105** (1996), 9299–9315.

[12] Ren, W. *Numerical Methods for the Study of Energy Landscapes and Rare Events*, thesis, New York University, 2002.

ICM 2002 · Vol. I · 631–646

International Comparisons in Mathematics Education: An Overview

Gabriele Kaiser* Frederick K. S. Leung†
Thomas Romberg‡ Ivan Yaschenko§

Abstract

The paper opens with an overview of the discussion of international comparisons (including goals) in mathematics education. Afterwards, the two most important recent international studies, the PISA Study and TIMSS-Repeat, are described. After a short description of the qualitative-quantitative debate, a qualitatively oriented small-scale study is described. The paper closes with reflection on the possibilities and limitations of such studies.

2000 Mathematics Subject Classification: 97.
Keywords and Phrases: International comparative studies, Mathematics education, Achievement studies, Qualitative case studies.

1. Goals of comparative studies

Since the results of the Third International Mathematics and Science Study (TIMSS) were published in 1996, international comparisons of student performance in mathematics have gained more and more importance as a consequence of public and political discussions. The discussions recently have been fueled by the results published in 2001 of the Programme for International Student Assessment (PISA). Nevertheless it has to be considered that comparative education has a long tradition going back to oral reports, as exemplified by Greeks and Romans. With the beginning of the 19th century, approaches were developed seeking to identify forces influencing the development of systems of education. In the 1960s and 1970s, the use of social science methods became common in order to examine the effect of

*University of Hamburg, Department of Education, Von-Melle-Park 8, 20146 Hamburg, Germany. E-mail: gkaiser@erzwiss.uni-hamburg.de
†Faculty of Education, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: hraslks@hku.hk
‡University of Wisconsin-Madison, Wisconsin Center for Education Research, 1025 West Johnson Street, Madison, Wisconsin, USA. E-mail: tromberg@facstaff.wisc.edu
§Moscow Center for Continuous Math Education, B. Vlas'evskij 11, 121002 Moscow, Russia. E-mail: ivan@mccme.ru

various factors on educational development accompanied by a debate on the relative merits of quantitative versus qualitative studies. We will come back to this discussion later on.

If we look for the goals of comparative education, history shows us that comparative education serves a variety of goals. It can deepen our understanding of our own education and society, be of assistance to policymakers and administrators, and be a valuable component of teacher education programmes. These contributions can be made through work that is primarily descriptive as well as through work that seeks to be analytic or explanatory, through work that is limited to just one or a few nations, and through work that relies on nonquantitative as well as quantitative data and methods. Based on that, Postlethwaite [11] discriminated four major aims of comparative education:

- "Identifying what is happening elsewhere that might help improve our own system of education" (p.xx). Postlethwaite gave several examples, such as the attempt to identify the principles involved in an innovation like mastery learning (which has had such success in the Republic of Korea) and grasping the procedures necessary to implement the mastery principle.
- "Describing similarities and differences in educational phenomena between systems of education and interpreting why these exist" (p.xx). This comprises the analysis of similarities and differences between systems of education in goals, in structures, in the scholastic achievement of age groups and so on, which could reveal important information about the systems being compared. Studies of these types might describe not only inputs to and processes within systems but also the philosophy of systems and outcomes. The reasons of why certain countries have particular philosophies and the implications these have in terms of educational outcomes are areas of both major academic and practical interest.
- "Estimating the relative effects of variables (thought to be determinants) on outcomes (both within and between systems of education)" (p.xx). Within education there is a great deal of speculation about what affects what. How much evidence, for example, do the people who teach methods at teacher-training establishments have on the effectiveness of the methods they promulgate? What about home versus school effects on outcomes? These questions and similar ones are the questions to be dealt with under this perspective.
- "Identifying general principles concerning educational effects" (p.xx). This means that we are aiming at a possible pattern of relationship between variables within an educational system and an outcome. In practice, a model will be postulated whereby certain variables are held constant before we examine the relationship between other variables and the outcomes. The resultant relationship will often be estimated by a regression coefficient. Principles we detect in an educational system that we analyze that recurs in other systems might be determined to be a general principle.

In mathematics education, there have been a remarkable number of international comparative studies carried out in the last 30 years. Robitaille [12, p. 41] believed that the reason for this might be that—

> Studies that cross national boundaries provide participating countries with a broader context within which to examine their own implicit theories, values and practices. As well, comparative studies provide an opportunity to examine a variety of teaching practices, curriculum goals and structures, school organisational patterns, and other arrangements for education that might not exist in a single jurisdiction.

Stigler and Perry [14, p. 199] emphasized the better understanding of one's own culture gained through comparative studies:

> Cross cultural comparison also leads researchers and educators to a more explicit understanding of their own implicit theories about how children learn mathematics. Without comparison, we tend not to question our own traditional teaching practices, and we may not even be aware of the choices we have made in constructing the educational process.

2. Recent international studies in mathematics

In the following, we will describe briefly the most important comparative studies in mathematics education (for details see [4]). Most of the large-scale studies have been carried out by the International Association for the Evaluation of Educational Achievement (IEA).

2.1. From FIMS over SIMS to TIMSS

The first large-scale international study was the First International Mathematics Study (FIMS), carried out 1964. Twelve countries participated in this study, in which two populations were tested—thirteen-year-olds and students in the final school year of the secondary school. In the first population, the students from Israel, Japan and Belgium received the best results, and the worst results were achieved by the U.S. students. In the second population, a different picture emerged—the youngsters from Israel, Belgium and England received the best results, and the U.S. students the worst. Several critics emphasized the important role of the curriculum and stated that valid comparative results cannot be formulated without considering curricular aspects.

The second large-scale comparative study in mathematics education was the Second International Mathematics Study (SIMS), 1980–1982, the results of which were published at the end of the 1980s and the beginning of the 1990s. Twenty countries participated in this study, which considered the same age groups as FIMS and contained a cross-sectional and a longitudinal component. Considering the curricular criticisms on FIMS, SIMS discriminated different levels of the curriculum—the intended curriculum, the implemented curriculum, and the attained curriculum. In addition, a content by cognitive-behavior grid was developed, which related the mathematical content with cognitive dimensions such as computation and comprehension. On the level of the intended curriculum, the main results were a significant curricular shift—geometry had lost importance in contrast to number and geometry.

On the level of the implemented curriculum, the study pointed out the different status of repetition in the different countries. On the level of the attained curriculum, the study showed that the increase in the achievements was remarkable low in many countries. Gender differences emerged in many countries, but were not consistent and were smaller than the differences between the different countries. SIMS has been criticised from several perspectives, and even the organisers of SIMS admitted that, despite the wealth of items, the curricula of many participating countries had not been covered sufficiently.

The last study in this series is the Third International Mathematics and Science Study (TIMSS), which was carried out in 1995 in over 40 countries. It examined the achievement of students from three populations at five grade levels (9-year-olds, 13-year-olds, and students in the final year of secondary school) in a wide range of content and performance areas, and it collected contextual information from students, teachers, and school principals. In considering the criticisms formulated at SIMS—the unsatisfactory coverage of the curriculum of the different countries, the focus on quantified outcomes (the quantified achievement of the students)—TIMSS established several additional studies:

- The TIMSS videotape study, which analyzed mathematics lessons in Japan, Germany, and the United States;
- The case study project, which collected qualitative information on the educational systems in Japan, Germany and the United States;
- The survey of mathematics and science opportunities, a study of mathematics and science teaching in six countries;
- The curriculum analysis study, which studied the curricula and textbooks in many countries.

With all these additional studies and the high number of countries participating in the main study, TIMSS remains the largest and most comprehensive study of educational practice in mathematics and science ever undertaken.

2.2. TIMSS repeat

Because of the impact of TIMSS on the international community, it was decided that a repeat study (TIMSS Repeat or TIMSS-R) be conducted so that trends in mathematics (and science) achievements could be studied in an international context. However, TIMSS was a complicated study involving testing three populations of students at five grade levels and in two subject areas. So it was decided that for TIMSS-R in 1999, only eighth grade students (i.e., the upper grade of the TIMSS population 2 level (i.e., 13-year-olds)) would be tested.

38 countries participated in TIMSS-R, and of the 38 countries, 26 had participated in the eighth grade test in TIMSS as well. So for these 26 countries, comparison between their eighth grade students' results in 1995 and 1999 could be made. 17 of these 26 countries had participated in the fourth grade test in TIMSS as well, and for these 17 countries, the choice of replicating the TIMSS study in 1999 means that the students tested in 1999 was the same cohort of students who took the TIMSS test in 1995. This thus constitutes a quasi-longitudinal study and

trends in achievement across the four-year duration can be studied. And for the 12 countries which did not participate in TIMSS in 1995, the TIMSS-R results would allow them to compare their students' achievements with all the TIMSS and TIMSS-R countries.

TIMSS-R, being a repeat study of TIMSS, adopted the TIMSS framework, which in turn was based on the SIMS framework. As pointed out above, SIMS and TIMSS placed special emphasis on the curriculum, and conceived the curriculum in the three levels of intended, implemented and attained curricula. The TIMSS curriculum framework, which was developed through a lengthy process of negotiation among National Research Coordinators of all the countries that participated in TIMSS with input from experts in the field of mathematics education, includes a Content dimension and a Performance expectations dimension (There is a third dimension known as Perspectives, but the TIMSS-R international report has not included results based on analysis of this dimension of the data). The five content areas tested included: Fractions and Number Sense (38% of the items were devoted to this area); Measurement (15%); Data Representation, Analysis and Probability (13%); Geometry (13%); and Algebra (22%). The categories under performance expectation were: Knowing (19%); Using Routine Procedures (23%); Using Complex Procedures (24%); Investigating and Problem Solving (31%); and Communicating and Reasoning (2%). The test consisted of items in multiple-choice and free-response (short answers and extended responses) formats, and about one quarter of the items and one third of the testing time were devoted to the free-response items.

Like TIMSS, questionnaires for school principals, teachers and students were administered to collect data on the variables related to student achievement. Also, in line with the IEA tradition, rigorous sampling and administration standards were established and closely monitored by the international study centre.

Results:

Mathematics Achievement

For the top performing countries in TIMSS-R, the pattern in TIMSS persisted. The East Asian countries (Singapore, Korea, Chinese Taipei, Hong Kong and Japan) outperformed their counter-parts in other parts of the world. Other countries that achieved well included Belgium (Flemish), Netherlands, Slovak Republic, Hungary, Canada, Slovenia, Russian Federation, Australia, Finland, Czech Republic, Malaysia and Bulgaria. Like TIMSS, one conspicuous finding in TIMSS-R is the magnitude of the difference in achievements among countries. The range of performance among countries is more than three standard deviations, and in fact, the achievements of all the five top performing East Asian countries were three standard deviations above that of the lowest scoring country, and those of the top six countries was more than two standard deviations above those of the three lowest performing countries.

For most of the countries that participated in both TIMSS and TIMSS-R, there was not much difference in terms of their relative performance in the two studies. Only one country had a significant decrease in its performance, and 3 out of the 26 had a significant improvement.

Variables Related to Achievement

System Variables

There was no clear relationship between the wealth of the countries (as measured by GNP) and their students' achievements. Although many affluent countries (Japan, Singapore, Belgium (Flemish), Netherlands, Hong Kong) did very well in TIMSS-R, some wealthy countries (such as US, Finland, Australia, Italy, and Israel) did not do as well. On the other hand, some less affluent countries (Slovak Republic, Korea and Chinese Taipei) did very well in TIMSS-R. Nor was high achievement related to public expenditures on education. In fact, the public expenditures on education of the top 9 countries were all less than the international average percentage of 5.13%.

For the average over all TIMSS-R countries, students with a higher level of educational resources at home and in school did better in the TIMSS-R test. However, we cannot conclude that students from countries with higher levels of educational resources did better in TIMSS-R. Some high achieving countries (such as Singapore and Hong Kong) had relatively few educational resources, while some countries with high level of resources (such as Israel and US) did not do very well. So while there might well be a positive correlation between educational resources and student achievement within countries, there did not seem to be a clear relationship between educational resources and achievement across countries.

Students' Attitudes towards Mathematics

Similar to the finding above, although the TIMSS-R results were consistent with the findings from the literature that students' positive attitudes towards mathematics was related with higher achievement within a country, the same relationship did not hold across countries. In fact, with the exception of Singapore, all the top-performing countries had relatively negative attitudes towards mathematics.

In particular, across countries, a positive self-concept in mathematics did not seem to be related with higher achievement. It is noticeable that students from all the five top-performing East Asian countries had very low self-image of mathematics. This suggests that self-image of mathematics or confidence in doing mathematics may be related to cultural values and is not necessarily associated with student achievement.

Teacher and Instructional Practices

The same pattern applied to teacher confidence. Although the TIMSS-R data show that within a country, teacher confidence was related with student achievement, teachers from high performing countries did not have particularly high level of confidence. In particular, although Japanese students did very well in TIMSS-R, their teachers had the lowest level of confidence among all the TIMSS-R countries.

Time devoted to mathematics instruction varied tremendously across countries, from the lowest of 73 hours and 9% of the total instructional time, to 222 hours and 17% of the total instructional time. But once again, there was no clear relationship between amount of instructional time and achievement. In actual fact, the four countries that devoted most time to mathematics instruction did badly in TIMSS-R, and the top performing countries were spending about just half of the time as these countries in instruction.

As far as instructional practices are concerned, teachers from participating countries reported that the two most predominant activities in their classrooms were teacher lecture and teacher-guided student practice, which accounted for nearly half of the class time. Although solving non-routine problems was mentioned in the intended curriculum of nearly all countries, teachers in all TIMSS-R countries, with the exception of Japan, reported that they put relatively low emphasis on mathematics reasoning and problem solving.

There were other discrepancies between the intended and the implemented curriculum. "For example, curricular goals and aims in 25 countries included "visualization of three-dimensional shapes" for all or almost all students, but teachers in only eight countries reported that at least 75 percent of the students had been taught this topic." ([9, p. 182])

Lastly, the amount of homework assigned by teachers to students differed tremendously across countries, but again there was no neat relationship between the amount of homework and the achievement of students.

Conclusion

As far as the TIMSS-R achievement within countries is concerned, the factors we discussed above merely confirmed the findings in previous studies, that achievement is related to educational resources at home and in school, with students' attitudes towards mathematics, with teachers' confidence etc. But what TIMSS-R failed to inform us is how to account for differences in performance across countries. As pointed out above, most of the variables that explained variation of achievement within countries failed to explain the variation across countries. This failure perhaps points to the limitation of questionnaires in getting at the factors for explaining student achievement. This problem is particularly acute in international studies, where the same term in a questionnaire may mean very different things in different culture. There is also the issue of confoundment between the cultural values and the measuring instrument. For example, the finding on negative self-concept in mathematics of East Asian students above may be due to the stress in the East Asian cultures of the virtue of humility or modesty. Children from these countries are taught from when they are young that one should not be boastful. This may inhibit students from rating themselves too highly on the question of whether they think they do well in mathematics, and so the scores may represent less than what students are really thinking about themselves. On the other hand, one's confidence and self image are something that is reinforced by one's learned values, and if students are constantly taught to rate themselves low, they may internalize the idea and may result in really low confidence.

For variables on instructional practices, the TIMSS-R teacher questionnaire results again did not give us a lot of clues on the instructional practices that will lead to high achievement. Probably, it is hard for instructional practices to be captured by self-reporting questionnaire, and that is why associated studies such as the TIMSS-R Video Study are so important in this regard. Video-taping offers a form of cross-cultural documentation which is both true to the original classroom and amenable to rigorous analysis [15], and is hence a much better methodology for studying instructional practices.

What TIMSS-R does tell us is that there exist vast differences in mathematics achievement across a large number of countries. Hopefully the realization of the differences will fuel a search for the factors that contribute to high achievement rather than a race to top the league table.

2.3. OECD/PISA

Developed jointly by the Organisation for Economic Cooperation and Development (OECD) member countries, the Program for International Student Assessment (PISA) is designed to monitor, on a regular basis, the literacy of students in reading, mathematics, and science as they approach the end of secondary school. The first PISA assessment took place in 2000 with the emphasis on reading, with initial assessments in both mathematics and science. The next assessment will be in 2003 with the emphasis on mathematics. PISA has been implemented through an international consortium led by the Australian Council for Educational Research (ACER) (for details see [10]).

The OECD/PISA Mathematical Literacy Study is concerned with the capacities of students to analyse, reason, and communicate ideas effectively as they pose, formulate, solve, and interpret solutions to mathematical problems in a variety of domains and situations. By focusing on real world problems, the PISA assessment does not limit itself to the kinds of situations and problems typically encountered in school classrooms. In real world settings, few people are faced with the drill-type of problems that typically appear in school textbooks and classrooms. Instead, citizens in every country are currently being bombarded with information on issues such as "global warming and the greenhouse effect," "population growth," "oil slicks and the seas," "the disappearing countryside," and so on, and a relevant question is whether citizens can make sense of claims and counterclaims on such issues. Of interest for the OECD/PISA study is whether 15-year-olds (the age when many students are completing their formal compulsory mathematics learning) can use the mathematics they have been taught to help make sense of these kinds of issues.

The concept of mathematical literacy, which underlies the OECD/PISA study is defined as—

> an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements, and to engage in mathematics, in ways that meet the needs of that individual's life as a constructive, concerned, and reflective citizen.

This definition of mathematical literacy is consistent with the broad and integrative theory about the structure and use of language as reflected in recent sociocultural literacy studies. The term "literacy" refers to the human use of language. In fact, each human language and each human use of language has both an intricate design tied in complex ways to a variety of functions. For a person to be literate in a language implies that the person knows many of the design resources of the language and is able to use those resources for several different social functions. Analogously considering mathematics as a language implies that students not only must learn the design features involved in mathematical discourse (the terms, facts,

signs and symbols, procedures, and skills in performing certain operations in specific mathematical subdomains and the structure of those ideas in each subdomain), they also must learn to use such ideas to solve nonroutine problems in a variety of situations defined in terms of social functions (making sense of some phenomena). Note that the design features for mathematics are more than knowing the basic terms, procedures, and concepts that one is commonly taught in schools. It involves how these features are structured and used. Unfortunately, one can know a good deal about the design features of mathematics without knowing either their structure or how to use those features to solve problems.

PISA assess mathematical literacy in three dimensions:

1. Content in terms of broad mathematics domains such as chance, change and growth, space and shape, uncertainty, among others. For PISA 2000, the mathematics assessment focused on two domains: change and growth, and space and shape.
2. Three "competency classes." The OECD/PISA mathematical literacy items have been developed to assess three classes of student mathematical competency:
   Class 1 competencies include those most commonly used on standardised assessments and classroom tests. These competencies are knowledge of facts, common problem representations, recognition of equivalents, recalling of familiar mathematical objects and properties, performance of routine procedures, application of standard algorithms and technical skills, manipulation of expressions containing symbols and formulae in standard form, and computations.
   Class 2 competencies include those related to students' planning for problem solving by drawing connections between the different mathematical content strands, or from different Big Ideas. They also include students' abilities to combine and integrate information in order to tackle and solve "standard" problems. Class 2 competencies reflect students' abilities to choose and develop strategies, to choose mathematical tools, to use multiple methods or steps in the mathematisation and modelling process. These competencies also reflect students' abilities to interpret and reflect on the meaning of a solution and the validity of their work. Problems that reflect student competencies in Class 2 require students to use appropriate elements from different mathematical content areas, or from different Big Ideas, in combination with conceptual thinking and reasoning based on material that does not call for large extensions of where the student has been before.
   Class 3 competencies include those related to students' ability to plan solution strategies and implement them in problem settings that are more complex and "original" (or unfamiliar) than those in Class 2. These competencies require students not only to mathematise more complex problems, but also to develop original solution models. Items measuring Class 3 competencies should reflect students' ability to analyse, to interpret, to reflect on, and to present mathematical generalisations, arguments and proofs.
3. Situations in which mathematics is used. For PISA, each item was set in one of five situation-types: personal, educational, occupational, public, and scien-

tific. The items selected for the mathematics test represent a spread across these situation types. In addition, items that can be regarded as authentic are preferred. That is, items should generally be mathematically interesting and should reflect problems that could be encountered through a person's day-to-day interactions with the world.

In a typical Competency Class 1 problem, students were asked to read information from a graph representing a physical relationship (speed and distance of a car). Students needed to identify one specified feature of the graph (the display of speed), to read directly from the graph a value that minimised the feature, and then select the best match from given alternatives.

For a Competency Class 2 problem, students were given a mathematical model (in the form of a diagram) and a written mathematical description of a real-world object (a pyramid-shaped roof) and asked to calculate the area of the base. The task required students to link a verbal description with an element of a diagram, to recall the area formula for a square with given sides, and to identify the required information in the diagram. Students then needed to carry out a simple calculation to find the required area.

A Competency Class 3 task required students to identify an appropriate strategy and method for estimating the area of an irregular and unfamiliar shape, and to select and apply the appropriate mathematical tools in an unfamiliar context. Students needed to choose a suitable shape or shapes with which to model the irregular area, know and apply the appropriate formulae for the shapes they used, work with scale, estimate length, and to carry out a computation involving a few shapes.

PISA 2000 results:

To summarise data from responses to a collection of such items, a five-level performance scale with an overall mean of 500 was created. The scale was created statistically using an Item Response Modelling approach to scaling ordered outcome data. Initially the overall scale was used to describe the nature of performance by classifying the nations in terms of overall performance, and thus to provide a frame of reference for international comparisons.

For PISA 2000, the rank-order of countries showed that 15-year-olds in Japan displayed the highest mean scores, but they could not be distinguished with statistical significance from scores in Korea or in New Zealand. Other countries that also scored above the OECD average were Australia, Austria, Belgium, Canada, Denmark, Finland, France, Iceland, Liechtenstein, Sweden, Switzerland, and United Kingdom. Overall, there was considerable within-country variation.

Of more importance was the relationship of other variables such as student motivation and engagement, gender, family background, and socioeconomic background to performance in mathematics:

- In most countries, because most 15-year-olds considered mathematics irrelevant to their future, only a small proportion considered mathematics worth pursuing.
- Lack of interest in mathematics was associated with poorer student performance.

- Males on average performed better than females on mathematical literacy, but the advantage disappeared when comparing low performers.
- Higher parental education and more social and cultural communication among parents and their children were associated with better student performance.
- Living with only one parent was, on average, associated with lower student performance.
- The socio-conomic composition of a school's student population was an even stronger predictor of student performance than individual home background.

In summary, the PISA 2000 results provided an interesting initial look at how 15-year-olds responded to a set of items constructed to assess mathematical literacy; the differences in mean performance across countries, and potentially important correlates of such performance.

3. The quantitative-qualitative debate and the case of a small-scale study

Since FIMS in 1964, it seems that the same questions have repeatedly been asked in large-scale studies, and that qualitative strategies are still not well considered, although as a result of the criticism of FIMS, SIMS was conceptualized as an in-depth-study of the curriculum. For the first time, issues such as those related to student and teacher beliefs were discussed. TIMSS added to SIMS studies, which aimed to explore the relationships between the intended, implemented, and attained curriculum.

For example, in the study Survey of mathematics and science opportunities [3], based on observations in mathematics and science in several countries, it was argued that there were typical patterns of instructional and learning activities in each country, which seemed to stem from the interaction of curriculum and pedagogy. It was assumed that students' learning experiences were moulded by teachers who selected, prepared and taught the mathematical content in a variety of instructional activities. In this respect, the researchers felt it necessary to elicit information on teachers' background knowledge, their beliefs about subject matter, and pedagogical beliefs. This view led the researchers to explore teachers' instructional practices in detail. As the description of teachers' practices through observations became the major focus, a major reorientation in paradigm and methodology was inspired. The orientation of conceptualization shifted from quantitative to qualitative differences between countries. The Case study project [13] was already designed at the beginning as an ethnographic study, aiming to combine large-scale surveys and qualitative methods. This in-depth studies of local situations intended to identify the myriad of causal variables that were not recognised in large scale surveys and to allow the development of hypotheses in order to interpret and explain many data gathered in large scale studies.

Apart from these qualitatively oriented case studies accompanying TIMSS, there exist many small scale international studies on mathematics education, and many more are coming. As an example of such studies, we briefly describe one

study comparing English and German mathematics teaching [5, 7]. This ethnographic study was carried out at the beginning of the 1990s, using methods of qualitative social sciences, mainly participating classroom observations. In general, the study aimed at generating general knowledge, based on which pedagogical phenomena might be interpreted and explained. Under a narrower perspective, the study aimed to generate hypotheses on the differences between teaching mathematics under the educational systems in England and in Germany. For methodological reasons, however, the study could not make any "lawlike" statements; in contrast, the study referred to the approach of the "ideal typus" developed by Max Weber (Webersche Idealtypen) and described idealized types of mathematics teaching reconstructed from the classroom observations in England and Germany. That means that typical aspects of mathematics teaching were reconstructed on the basis of the whole qualitative studies rather than on an existing empirical case.

In brief, the study concluded that the following general approaches concerning the understanding of mathematical theory were predominant in English and German mathematics education. In German mathematics teaching, a subject-oriented understanding of mathematical theory prevailed, in contrast to the prevalence of a pragmatic understanding of mathematical theory in England. Generally speaking, mathematics teaching in Germany was characterized by its focus on the subject structure of mathematics and on mathematical theory. This meant that theory was made explicit by means of rules and computations. In contrast, in England, the understanding of theory could be called pragmatic—theory was applied practically in an appropriate way. These different basic approaches in England and Germany to teaching mathematics were visible when looking at the differences with regard to the following aspects.

The focus on theory when teaching mathematics in Germany implied a lesson structure which went along with the subject structure of mathematics. Thus, in the lessons, large units were complete in themselves. Mathematical theorems, rules and formulae were therefore of high importance. That varied, though, with the different kinds of schools of the three-track system.

The approach in England, the pragmatic understanding of theory, was apparent from the curricular structure, which resembled a spiral. As a consequence, smaller units were taught, and they were not necessarily connected with each other. Topics were quickly swapped, and at times different topics are worked on at the same time. Frequent repetitions of mathematical terms and methods that had already been taught were a feature of this spiral-shaped approach. Mathematical theorems, rules, and formulae (often called "patterns") were of low importance for the teaching of mathematics in England.

In Germany, proofs of mathematical statements were to a certain extent important when teaching mathematics at the schools of the higher achievement level, but they had only small or nearly no importance in schools of the intermediate or lower achievement level. Proofs were considered important in order to visualise the theoretical frame of mathematics, especially in the context of geometry.

In England, proofs were of low importance, both in selective as well as in nonselective schools. Theorems, found by means of experiments, were often only

checked with examples, and proofs and checks with examples were often not distinguished.

The status and role of proofs in Germany was studied more extensively in another qualitative small-scale study by Knipping [8], comparing the proof of the theorem of Pythagoras in French and German mathematics classes.

4. Strengths and limitations of international studies in mathematics

The strengths of comparative education can be seen in the multidimensional aims described at the beginning such as describing similarities and differences in educational phenomena in different educational systems, estimating the effect of special variables on the outcomes, based on input-output-models of education, identifying general principles concerning educational effects. One major stream of comparative work concerns itself with the interaction of educational and political, social, or economic systems, for which the PISA study is an example providing politicians with educational indicators, which aim to steer educational systems. Another stream focuses on particular pedagogical factors, for which the different case studies accompanying TIMSS are an example. Comparisons of instructional methods, curricula, teacher training, and their presumed outcomes (student behavior, especially achievement) have long been at the heart of comparative work, although the focus of attention has recently broadened. Alexander [1, p. 149] describes this broadened view as follows:

> I argue that educational activity which we call pedagogy—the purposive mix of educational values and principles in action, of planning, content, strategy and technique, of learning and assessment, and of relationships both instrumental and affective—is a window on the culture of which it is a part, and on that culture's underlying tensions and contradictions as well as its publicly-declared educational policies and purposes.

On the other hand, the limitations or even dangers of international comparative studies are also now widely discussed among the researchers.

Kaiser [6] described the following alternative approaches in comparative education, which challenged established research traditions in comparative education since the 1980s. The first challenge to the nation-state as the exclusive research framework either looked at the world system—regional variations, racial groups, classes, which are not bound to the nation—or did microanalytic research focused on regional variation. Proponents of the analysis of regional variation argued that educational variance often was as great, if not greater, across regions within a nation as it was across nations.

The second challenge to input-output models and total reliance on quantification was based on assertions that education and school practices could not be reduced solely to quantitative aspects—knowledge about these topics could only be generated by qualitative research methods that focused on actual, lived educational practices and processes. A few approaches proposed ethnomethodological

techniques and related educational processes to broader theories of school-society relations.

The third challenge questioned the dominance of structural functionalism in comparative education—either how education functioned to maintain the social fabric, or how it could be made to function (in the case of the Third World) to develop a nation-state generally along Western models. New approaches proposed conflict theories, because most societies are plural societies characterised by conflict, in which dominant groups seek to legitimise their control over the state.

The fourth challenge, the emergence of new research concerns, involved new ways of looking at educational institutions and their relation to society—studies on the nature of knowledge transfer and its impact on the Third World, on school knowledge, and on the internal workings of the school.

Concerning the international comparisons in mathematics education Kaiser [6] described four critical aspects, which should be considered in the coming debate:

- On the methodological level, it is necessary to ask whether the main concepts and ideas of the methodology used, such as the approach of the probabilistic test theory, are adequate. This model delivers highly general results about ability levels. It is, on the other hand, worth questioning whether the necessary general conditions required by the model—existence of a one-dimensional construct "mathematical ability" or "mathematical literacy"—are fulfilled.
- On a more subject-bound level, curricular issues have to be discussed—the dependence on the results from the test items, the adequacy of the test items concerning so many different curricula, the relation of achievement results, and the opportunity to learn.
- Under a general pedagogical perspective, there is a question of the innovation potential of such studies. What can we learn from descriptions of mathematics teaching in totally different cultures, such as Germany and Japan, with very different value systems, social conditions, and so on?
- Under a political perspective, we need to ask whether the scientific community is able to control how the results of the studies are used in political debates. Consensus exists among researchers that the ranking is not the main result of such studies—but the public debate concentrates mainly on the ranking lists and bases proposals for consequences on the rank achieved. How far are the results of such studies under the control of the researchers involved, and how far can researchers influence the usage of the results?

Other branches of criticism emphasize the influence of multiple-choice tests on mathematics education and on the promotion of mathematically talented youngster: We have to distinguish between the mathematical education for youngsters, who are highly interested and talented in mathematics and intend to become a mathematician, engineer, etc. and mathematics education for the others, who will not be necessarily involved in vocations dealing with mathematics. The goals and the methods of those two different branches of mathematical education are quite different and should therefore use different methods of comparison.

In many countries of the world, there is an elaborate system of mathematical olympiads. They help to popularize mathematics, to engage talented students

into mathematics. The results of students in major mathematical competitions (such as IMO) are sometimes used to compare the levels of education in different countries. The results of the team in IMO might be one indicator for the level of mathematical education in a country. In addition, there is another extraordinary international competition—the Tournament of the Towns, —which has a stronger focus on broader groups of mathematically interested and talented students. It offers an opportunity for students of many countries to compete with other students in solving ambitious interesting mathematical problems.

Coming back to the already mentioned large-scale international studies, whose main part of the items are still multiple-choice items. Those items are in many respects not adequate to compare the achievements of mathematically talented students: One of the main ideas of classical mathematical education looking for those students assumes that a student is being taught not only to find solutions for different kinds of problems, but to create different ways of solving problems, to create arguments for the solution and communicate it orally and in written form to an audience. Such a student is used to have a relatively large amount of time per problem and he/she will have significant difficulties in solving multiple-choice items under time pressure, where no sophisticated argumentation is asked.

We have to consider that mathematics education in different countries is very sensible to means of testing. Thus bringing multiple-choice tests into education often leads to harmful changes in the whole education process. In Russia for example, it has led to a degradation of an ambitious mathematics education in many schools. It is an open question so far, whether mathematical olympiads or other tournaments might be an appropriate indicator for the level of mathematical education achieved by the students of a country, at least by the mathematically talented, which could to a certain extent replace large-scale international comparisons.

To summarize, comparative education is characterized by a wide diversity of approaches, perspectives, and orientations, and this diversity of the field seems to be one of its main strengths.

References

[1] Alexander, R., *Culture in Pedagogy, Pedagogy across Cultures*, Alexander, R. & Broadfoot, P. & Philipps, D. (Eds), Learning from Comparing: New Directions in Comparative Educational Research. Vol. 1, Symposium Books, Oxford, 1999, 149–180.

[2] Beaton, A. et al., *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*, Boston, Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1996.

[3] Cogan, L.S. & Schmidt, W.H., *An Examination of Instructional Practices in Six Countries*, [4], 48–67.

[4] Kaiser, G. & Luna, E. & Huntley, I. (Eds), *International Comparisons in Mathematics Education*, Falmer Press, London, 1999.

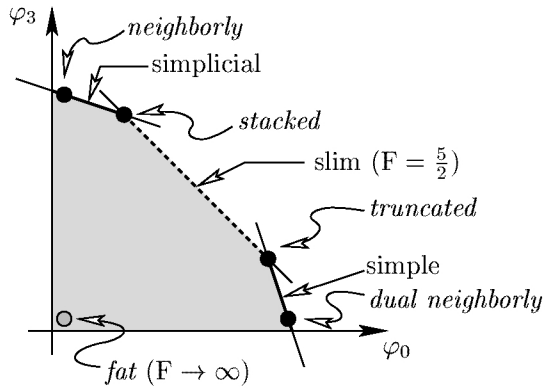
[5] Kaiser, G., *Comparative Studies on Teaching Mathematics in England and Germany*, [4], 140–150.
[6] Kaiser, G., *International Comparisons in Mathematics Education Under the Perspective of Comparative Education*, [4], 3–15.
[7] Kaiser, G., *Unterrichtswirklichkeit in England und Deutschland. Vergleichende Untersuchungen am Beispiel des Mathematikunterrichts*, Weinheim, Studien Verlag, 1999.
[8] Knipping, C., *Proof and Proving Processes: Teaching Geometry in France and Germany*, Peter-Koop, A. et al. (Eds), Developments in Mathematics Education in German-speaking Countries, Vol. 4, Franzbecker, Hildesheim, 2000, 44–55.
[9] Mullis, I.V.S. et al., *TIMSS 1999 International Mathematics Report*, International Study Center, Lynch School of Education, Boston College, Boston, 2000.
[10] Organisation for Economic Co-operation and Development (Ed), *Knowledge and Skills for Life. First Results from PISA 2000*, Paris, OECD, 2001.
[11] Postlethwaite, T.N., *The Encyclopeadia of Comparative Education and National Systems of Education, Preface*, Oxford, Pergamon, 1988, xvii–xxvii.
[12] Robitaille, D.F., *TIMSS: The Third International Mathematics and Science Study*, Beitraege zum Mathematikunterricht, Hildesheim, Franzbecker, 1994, 35–42.
[13] Stevenson, H.W., *The Case Study Project of TIMSS*, [4], 104–120.
[14] Stigler, J.W. & Perry, M., *Cross Cultural Studies of Mathematics Teaching and Learning: Recent Findings and New Directions*, Grouws, D.A. & Cooney, T.J. & Jones, D. & (Eds), Perspectives on Research on Effective Mathematics Teaching, Reston, National Council of Teachers of Mathematics, 1988, 194–223.
[15] Stigler, J.W. et al., *The TIMSS Videotape Classroom Study, Office of Educational Research and Improvement*, U.S. Department of Education, Washington, 1999.

Appendix B: Errata and Author's Modifications for Volumes II and III

Errata for Volumes II and III

page	line	instead of	read
II-72	8 from top	in §	in §2
II-152	18 from top	in 2.	in 2.1
II-154	4 from top	in 2.	in 2.1
II-154	5 from top	in 3.	in 3.1
II-154	6 from top	(cf. section 2.)	(cf. section 2.4)
II-156	18 from top	(see 2.)	(see 2.4)
II-157	22 from top	in section 4.	in section 4.2
III-313	9 from top	of section	of section 2
III-313	10 from bottom	see again section	see again section 2
III-314	22 from top	in section	in section 1
III-766	2 from bottom	to section	to section 6

III-630: replace the figure as follows



Author's Modifications in Volumes II and III

Norm Varieties and Algebraic Cobordism

Markus Rost

Page 77 in Volume II: add **2000 Mathematics Subject Classification** 12G05.

Page 77: add **Keywords and Phrases** Milnor's K-ring, Galois cohomology, Cobordism.

Line 1 from top on page 81: replace "$[1, \zeta^i]$" by "$[1, \zeta^i \sqrt[p]{c}]$".

Line 13 from top on page 81: replace "t^p" by "ct^p".

Knots, von Neumann Signatures, and Grope Cobordism

Peter Teichner

Page 443 in Volume II: replace Figure 2 as follows

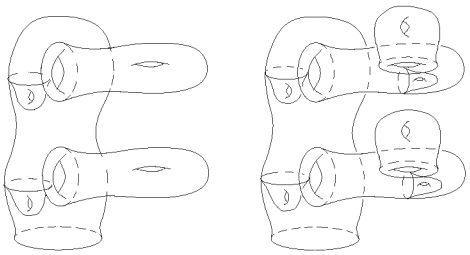


Figure 2: Symmetric gropes of height 2 and 2.5

On the dynamical Yang-Baxter equation

Pavel Etingof

Line 21 from top on page 559 in Volume II: replace "[Fe, EV2, FV1, TV]" by "[Fe, EV2, FV1, TV1]".

Line 6 from top on page 560: replace "is known in the non-dynamical case ([EK]), and was proved" by "is proved in [EK] in the non-dynamical case, and".

Line 15 from bottom on page 560: replace "$\to$" by "$\mapsto$".

Line 7 from top on page 566: replace "$R_{VW}(z, \hat{\lambda})$" by "$R_{VW}(u, \hat{\lambda})$".

Line 11 from top on page 566: add "For $\mathfrak{g} = \mathfrak{sl}_n$, they were computed in [TV2]." in the end of paragraph.

Line 2 from top on page 567: replace "in [TV]" by "in [TV1]".

Line 11 from top on page 569: replace "(see [EV5])" by "(see [TV3, EV5])".

Line 3 from top on page 570: replace Ref. [EV5] by "[EV5] Etingof P., Varchenko A., *Dynamical Weyl groups and applications*, math.QA/0011001."

Line 7 from top on page 570: replace Ref. [Fe] by "[Fe] Felder G., *Conformal field theory and integrable systems associated to elliptic curves*, Proceedings of the International Congress of Mathematicians, Zürich 1994, p.1247–1255."

Line 4 from bottom on page 570: replace "[TV]" by "[TV1]", and add the following two references:

[TV2] Tarasov, V., Varchenko, A., Geometry of q-hypergeometric functions, quantum affine algebras and elliptic quantum groups. Astrisque No. 246 (1997).

[TV3] Tarasov, V., Varchenko, A., Difference equations compatible with trigonometric KZ differential equations. IMRN 2000, no. 15, 801–829.

Bubbling and regularity issues in geometric non-linear analysis

T. Rivière

Replace "*no-neck property*" by "*energy identity*" in the whole manuscript.

Line 18 from top on page 198 in Volume III: replace "ε_0" by "$\varepsilon(m, N)$".

Line 19 from bottom on page 198: replace "$\epsilon-$regularity results has been found in various problems of geometric analysis (minimal surfaces, geometric flows, Yang-Mills Fields...etc)." by "and previous partial regularity result of that type for minimal surfaces and other elliptic systems by Almgreen, DeGiorgi, Giusti, Miranda, Morrey, $\cdots$, $\epsilon-$regularity theorems has been found in various problems of geometric analysis (geometric flows, Yang-Mills Fields, $\cdots$, etc)."

Line 2 from top on page 200: replace "(see [Pa])" by "(see the work of Parker [Pa] that was following his bubbling picture he established in collaboration with J.Wolfson for pseudo-holomorphic curves)."

Line 21 from top on page 202: delete "It is even unknown, in the general case, whether the limiting map u is weakly harmonic on the whole of B^m."

Line 26 from top on page 202: add "Their proof is based on an ϵ-regularity

result for weak Yang-Mills in high dimension which was also independently obtained by Y. Meyer and the author in [MR]." in the end of paragraph.

Line 19 from top on page 203: delete "Question ii) is studied in a work in progress with F. Pacard."

Line 24 from top on page 207: add the following reference:
[MR] Y. Meyer and T. Rivière, Partial regularity for a class of stationary Yang-Mills Fields, to appear in Rev. Math. Iberoamericana, (2002).

Non uniformly hyperbolic dynamics: Hénon maps and related dynamical systems

Michael Benedicks

Line 2 from top on page 260 in Volume III: replace "$\|D(\log|\det Df|)\|_{C^1}$" by "$\|D(\log|\det Df|)\|_\infty$".

Action Minimizing Solutions of the Newtonian n-body Problem: From Homology to Symmetry

A. Chenciner

Line 1 from bottom on page 280 in Volume III: add "half" in the front of the sentence "its time derivative $\cdots$".

Line 6 from top on page 281: replace "$|\vec{r}_j(t)-\vec{r}_i(t)|^3$ by "$\|\vec{r}_j(t)-\vec{r}_i(t)\|^3$".

Line 7 from bottom on page 281: add "existence is not a problem here because fixing the ends gives coercivity" in the end of the sentence "$\cdots$ (see [C3]".

Line 4 from top on page 282: replace "in a minimizer" by "in a (local) minimizer".

Line 13 from top on page 282: replace "but no proof appeared" by "but no proof has yet appeared".

Line 7 from bottom on page 283: replace "enters" by "leaves".

Line 2 from top on page 284: delete "and [Ma]".

Line 4 from top on page 284: replace "coïncide" by "coincide".

Line 12 from bottom on page 287: replace "Venturelli's" by "The", replace "[V1]" by "[V1], [ZZ1]".

Line 11 from bottom on page 287: replace "homogy" by "homology".

Line 4 from top on page 288: replace "the bodies" by "equal-mass bodies".

Lines 5 and 7 from top on page 288: replace "fondamental" by "fundamental".

Line 8 from top on page 288: replace "minimizes in all cases where" by "is action minimizing when".

Line 14 from bottom on page 288: replace "recall [AC] that" by "recall that, according to [AC],"

Replace Paragraph 3 from bottom on page 288 by

(iii) Eights with less symmetry. As another example, we prove the existence, for three equal masses, of solutions "of the Eight type" but with less a priori symmetry than the full symmetry group $D_6 = \{s, \sigma | s^6 = 1, \sigma^2 = 1, s\sigma = \sigma s^{-1}\}$ (see [C3]) of the space of oriented triangles ("shape sphere" in [CM]). We consider the subgroups $\mathbb{Z}/6\mathbb{Z} = \{s\}$ and $D_3 = \{s^2, \sigma\}$.

Line 14 from top on page 289: replace "minimizes" by "is supposed to minimize".

Line 8 from bottom on page 289: delete "For $\pi/6 \leq u \leq \pi/3$, the minimizer is a horizontal Lagrange solution whose size increases to infinity and action decreases to 0. The x^4-type bifurcation of the minimizer at $u = \pi/6$ was analyzed by Marchal."

Line 16 from top on page 291: replace "Simo" by "Simó".

Line 9 from top on page 294: add the following two references:

[ZZ1] Zhang S. & Zhou Q., A Minimizing Property of Lagrangian Solutions, *Acta Mathematica Sinica, English Series*, Vol. 17, No.3 (2001), 497–500.

[ZZ2] Zhang S. & Zhou Q., Variational method for the choreography solution to the three-body problem, *Science in China (Series A)*, Vol. 45 No. 5 (2002), 594–597.

Replace paragraph 2 from top on page 290 to paragraph 2 from top on page 291 by

Thanks to the fact that a minimizer of the fixed-ends problem has no collision, we need only show that a minimizing path has no collisions at its ends, which can be done using local deformations as in one of the proofs of the test assertion for the Kepler problem. The surprise is that, using as a model the horizontal Lagrange family (which satisfies the symmetry requirements), one can give a simple direct proof of the absence of collisions in a minimizer:

1) the action of an admissible path undergoing a collision is bigger than the action $\hat{A}_2 = 2^{-\frac{5}{3}} 3^{\frac{2}{3}} \pi^{\frac{2}{3}} T^{\frac{1}{3}}$ (masses =1) of the horizontal relative equilibrium solution x_0 of an equilateral triangle which rotates by $\frac{\pi}{3}$ in the same amount of time $T/12$;

2) this last action is, for any $u \leq \frac{\pi}{3}$, bigger than the one $A(u) = \hat{A}_2 \left[\frac{3}{\pi}(\frac{\pi}{3} - u)\right]^{\frac{2}{3}}$ of the horizontal relative equilibrium solution x_u of an equilateral triangle which rotates by an angle $(\frac{\pi}{3} - u)$ during the given amount of time.

The first estimation, better than the one in [CM] ($\hat{A}_2 = 2^{\frac{1}{3}} A_2$) , appears in

[ZZ2] in the case of the Eight. It follows from the remark (at the basis of [V1] and [ZZ1]) that the action of a 3-body problem splits into the sum of three terms, each of which is one third of the action of the Kepler problem with attraction constant equal to the total mass $M = 3$. As the configurations at t and $t+T/2$ are symmetric with respect to the horizontal plane (compute $\alpha(s^3)$ and $\beta(s^3)$), any collision which occurs at t_0 occurs also at $t_0 + T/2$. The lower bound of the Kepler action during a period T is then twice the minimum of the Kepler action of an ejection-collision wih attraction constant 3 and period $T/2$. But this is exactly the action of x_0 during the period.

Finally, we prove that, for $0 \leq u < \frac{\pi}{6}$, the Lagrange solution x_u is not a minimizer. This is because the value $d^2\mathcal{A}(x_u)(\xi,\xi)$ of the Hessian of the action on the vertical variation

$$\xi = \left(\sin(\frac{2\pi t}{T}),\ \sin(\frac{2\pi t}{T} + \frac{2\pi}{3}),\ \sin(\frac{2\pi t}{T} + \frac{4\pi}{3})\right)$$

which "opens" x_u in the direction of the Eight, is negative for $u < \pi/6$ and positive for $u > \pi/6$. Indeed, the Hessian of x_u is positive when $\pi/6 < u \leq \pi/3$, which supports Marchal's claim that x_u is the minimizer when $\pi/6 \leq u \leq \pi/3$ (notice that its size increases to infinity and its action decreases to 0 when u tends to $\pi/3$).
Questions. 1) Prove that for $u = 0$, (the) minimum is planar, hence (the) Eight. 2) Our argument works for one value of u at a time. As no uniqueness is proved, neither is continuity with u of the family. Such continuity would imply the existence among the family of spatial 3-body choreographies in the fixed frame.
Remark. The first continuation of the Eight into a family of rotating planar choreographies was given by Michel Hénon [CGMS] using the same program as in [H]. A third family should exist, rotating around an axis orthogonal to the first two.

Mirror Symmetry and Quantum Geometry

Kentaro Hori

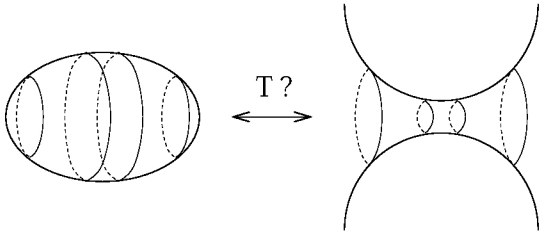


Figure 1: Is this what T-duality does?

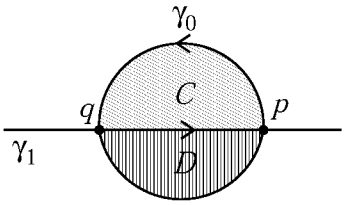


Figure 2: An example with $Q^2 \neq 0$

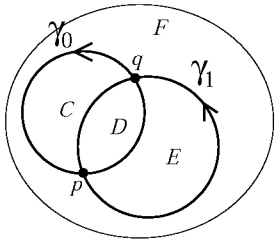


Figure 3: Intersecting Lagrangians in S^2

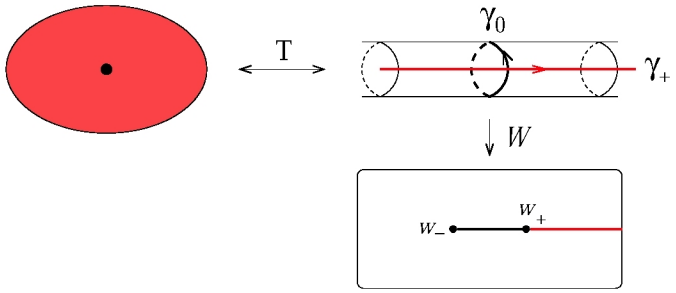


Figure 4:

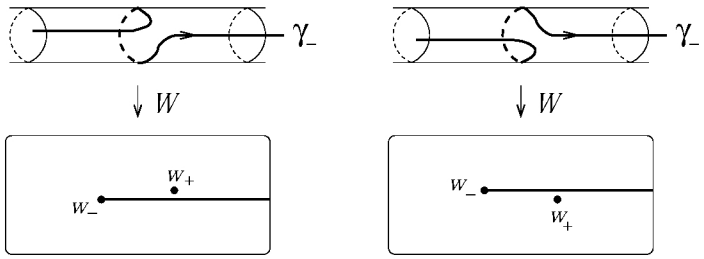


Figure 5:

Author Index for Volumes I, II and III