

PROCEEDINGS OF THE
INTERNATIONAL CONGRESS OF
MATHEMATICIANS

AUGUST 3–11, 1986



INTERNATIONAL CONGRESS OF MATHEMATICIANS

Berkeley, California

August 3-11, 1986

EDITED BY

ANDREW M. GLEASON

Library of Congress-in-Publication Data

International Congress of Mathematicians (1986: Berkeley, Calif.)

Proceedings of the International Congress of Mathematicians, 1986: August 3-11, 1986, Berkeley, California, USA.

Includes index.

1. Mathematics--Congresses. I. American Mathematical Society. II. Title.

QA1.I82 1986 510 87-24109

ISBN 0-8218-0110-4

The 1986 International Congress of Mathematicians was supported in part by Grant INT-8417900 from the National Science Foundation including contributions from the Department of the Air Force, Air Force Office of Scientific Research, and the Department of Energy, Office of Energy Research. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Department of the Air Force, or the Department of Energy.

This work relates to Department of Navy Grant N00014-85-G-0145 issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein.

The 1986 International Congress of Mathematicians was supported in part by Grant DAAG29-85-G-0091 from the Department of the Army. The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

Photo credits: William G. Chinn, Benedict H. Gross, Hikosaburo Komatsu, Jill P. Mesirov, and Klaus Peters.

Copyright © 1987 by the American Mathematical Society.

Printed in the United States of America.

All rights reserved except those granted to the United States Government.

The paper used in this book is acid-free and falls
within the guidelines established to insure permanence and durability. ☹

PREFACE

The *Proceedings* of the International Congress of Mathematicians, 1986, are printed in two volumes.

Volume 1 contains the official record of the Congress held at Berkeley from August 3 to August 11, 1986, the list of Ordinary Members, the reports on the work of the Fields Medalists and the Nevanlinna Prize Winner, the Plenary addresses, and the addresses in sections 1–8. Volume 2 contains the addresses in sections 9–19.

The Short Communications made by members at the Congress are not included in the *Proceedings*, but the names of the communicators and the titles of their papers will be found by section in the scientific program. Summaries of these contributions, if received in time, were printed in *Abstracts*, issued to members during the Congress.

Berkeley, 1986

CONTENTS

PAST CONGRESSES	VII
PAST FIELDS MEDALISTS AND NEVANLINNA PRIZE WINNERS	IX
ORGANIZATION OF THE CONGRESS	XI
THE CONGRESS AND ITS COMMITTEE STRUCTURE	XIII
LIST OF DONORS	XVII
OPENING CEREMONIES	XXI
CLOSING CEREMONIES	XXXV
SCIENTIFIC PROGRAM	XLIII
CONGRESS PARTICIPANTS	LXXVII
MEMBERSHIP BY COUNTRY	CI
THE WORK OF THE FIELDS MEDALISTS	1
INVITED ADDRESSES AT THE PLENARY SESSIONS	23
INVITED ADDRESSES AT THE SECTION MEETINGS:	
SECTIONS 1–8 (VOLUME 1)	305
SECTIONS 9–19 (VOLUME 2)	871
AUTHOR INDEX	A 1

PAST CONGRESSES

1897 Zurich
1900 Paris
1904 Heidelberg
1908 Rome
1912 Cambridge, U.K.
1920 Strasbourg
1924 Toronto
1928 Bologna
1932 Zurich
1936 Oslo

1950 Cambridge, U.S.A.
1954 Amsterdam
1958 Edinburgh
1962 Stockholm
1966 Moscow
1970 Nice
1974 Vancouver
1978 Helsinki
1982 Warsaw (held in 1983)



**Ahlfors: 1936 Fields Medalist,
Honorary President of the Congress**

PAST FIELDS MEDALISTS AND NEVANLINNA PRIZE WINNERS

Recipients of Fields Medals

1936 Lars Ahlfors	1970 Alan Baker
Jesse Douglas	Heisuke Hironaka
1950 Atle Selberg	Sergei Novikov
Laurent Schwartz	John Thompson
1954 Kunihiko Kodaira	1974 Enrico Bombieri
Jean-Pierre Serre	David Mumford
1958 Klaus Roth	1978 Pierre Deligne
René Thom	Charles Fefferman
1962 Lars Hörmander	Gregori Margulis
John Milnor	Daniel Quillen
1966 Michael Atiyah	1982 Alain Connes
Paul J. Cohen	William Thurston
Alexander Grothendieck	Shing-Tung Yau
Stephen Smale	

Nevanlinna Prize Winner

1982 Robert Tarjan



Lehto, Moser, Mesirov

ORGANIZATION OF THE CONGRESS

The National Academy of Sciences of the United States, through the U.S. National Committee for Mathematics, invited the International Mathematical Union to hold the 1986 International Congress of Mathematicians in the United States at the University of California, Berkeley. The invitation was accepted by the International Mathematical Union at the Warsaw Congress in August, 1983. Final action was taken a few days later when the Warsaw Congress at its Closing Ceremonies accepted the invitation to Berkeley, which was presented by Professor Jack K. Hale on behalf of the mathematical community of the United States of America.

The Academy, in turn, asked the American Mathematical Society to handle the organizational aspects of the Congress. The Society organized the Congress as a nonprofit corporation, ICM-86, and Jill P. Mesirov was appointed Executive Director. ICM-86 used the services of the American Mathematical Society's Meetings Department headed by H. Hope Daly, who served as Congress Manager.

The scientific program was organized by a Program Committee appointed by the International Mathematical Union. Its members were Enrico Bombieri, Lennart Carleson, Friedrich E. P. Hirzebruch (Chairman), David Mumford, Louis Nirenberg, Michael O. Rabin, J. A. Rozanov, David P. Ruelle, and I. M. Singer. The committee divided the program of the Congress into 19 sections and appointed a panel to nominate speakers in each section. The committee also approved the inclusion of sessions for ten-minute Short Communications in the Congress. In October of 1985 the Program Committee selected 16 mathematicians to give Plenary Addresses and 148 to give lectures in sections. Invitations were issued over the following weeks. Of the 164 invited, 14 plenary and 132 section speakers were present at the Congress.

The Fields Medals Committee, consisting of P. Deligne, J. Glimm, L. Hörmander, K. Ito, J. Milnor, J. Moser (Chairman), S. Novikov, and C. S. Seshadri arrived at its decisions in early 1986. The Nevanlinna Prize Committee, whose members were S. Cook, L. D. Faddeev (Chairman), and S. Winograd, also completed its work early that year.

The Board of Directors of ICM-86 appointed a Steering Committee, chaired by Andrew M. Gleason of Harvard University, to oversee the arrangements. There were several subcommittees of the Steering Committee. The Local Arrangements Committee was chaired by John W. Addison Jr. of the University of California, Berkeley; the Scheduling and Abstracts Committee consisted of Kenneth A. Ross of the University of Oregon and Hugo Rossi of the University of Utah; the Committee on Special Funds was chaired by Richard D. Anderson, Louisiana State University; and the Public Information Committee was chaired by Yousef Alavi of Western Michigan University.

Late in 1984, preliminary announcements of the Congress were sent out to all countries where some mathematical organization could be located, requesting information on the number of copies of the First Announcement they required. The First Announcement, prepared in English, French, German, and Russian,

was sent to these mathematical organizations in early 1985. The Second Announcement contained detailed information about the Congress, instructions about submitting abstracts for Short Communications, and the preregistration and housing form. It was sent to over 8,000 individuals starting in December of 1985. The Third Announcement, which contained a preliminary version of the scientific program, was distributed with the acknowledgment of preregistration between March and June of 1986. In all, 3,586 Ordinary Members, 340 Accompanying Members, and 69 Child Members registered for the Congress. Of these 721 Ordinary Members, 24 Accompanying Members, and 7 Child Members registered on site.

The Congress had several sources of revenue. A subvention was received early on from the International Mathematical Union. Grants were later received from the Air Force Office of Scientific Research, Army Research Office, Office of Energy Research, Office of Naval Research, National Science Foundation, Alfred P. Sloan Foundation, and the Vaughn Foundation Fund. Through the efforts of the staff and the Committee on Special Funds, donations in the form of cash, goods, and services were received from a number of both public and private individuals and corporations. It is especially interesting to note that nearly 9,000 members of the American Mathematical Society made contributions to the Congress totaling over \$30,000. Registration fees were \$125 for Ordinary Members registered before May 15, 1986, and \$163 thereafter; \$60 and \$78 for Accompanying Members; and \$30 for Child Members.

The International Mathematical Union gave travel grants to 32 young mathematicians from developing countries or from countries with currency difficulties. The Congress waived their registration fees, and, thanks to funds received from the Sloan Foundation and GTE Laboratories Incorporated, was able to give them room and board in the residence halls.

All sessions of the Congress took place on the campus of the University of California, Berkeley. The Opening Ceremonies were held in the Greek Theater. All other plenary sessions were held in Zellerbach Auditorium. These sessions were simultaneously broadcast over closed circuit television to several large lecture halls. Video tapes of the plenary lectures were shown in the evenings. The lectures in the sections were given at a number of locations on the campus.

In addition to the invited lectures, about 700 ten-minute Short Communications were given and a large number of informal seminars were arranged by participating mathematicians. Abstracts of the Short Communications were published in a book distributed to all Ordinary Members. An educational materials exhibition consisting of 40 booths representing 26 exhibiting firms was open throughout the Congress.

The Steering Committee arranged various social events. The Chancellor of the University of California, Berkeley hosted an outdoor reception in the Faculty Glade on the first evening of the Congress, Sunday, August 3. A western-style barbecue and rodeo was held Thursday evening, August 7, in the Cow Palace in San Francisco, where 2,800 members attended. A jazz concert was presented on Saturday, August 9, and a classical concert on Sunday, August 10. Each of these concerts was attended by approximately 1,500 participants.

THE CONGRESS AND ITS COMMITTEE STRUCTURE

BOARD OF DIRECTORS

Steve Armentrout, Pennsylvania State University
Frederick W. Gehring, University of Michigan
Chairman: Ronald L. Graham, Mathematical Sciences Research Center,
AT&T Bell Laboratories
Secretary: Jill P. Mesirov, Thinking Machines Corporation
George D. Mostow, Yale University (Member, IMU Executive
Committee)
Treasurer: Franklin P. Peterson, Massachusetts Institute of Technology

STEERING COMMITTEE

John W. Addison, Jr., University of California, Berkeley
Yousef Alavi, Western Michigan University
Peter J. Bickel, University of California, Berkeley
Hirsh G. Cohen, Research Division, IBM
H. Hope Daly, American Mathematical Society
Solomon Feferman, Stanford University
Chairman: Andrew M. Gleason, Harvard University
Leon A. Henkin, University of California, Berkeley
Shirley A. Hill, University of Missouri
Richard M. Karp, University of California, Berkeley
Linda Keen, Lehman College, CUNY
Jill P. Mesirov, Thinking Machines Corporation
Henry O. Pollak, Mathematical Communication & Computer
Sciences Research Laboratory, Bell Communications
Research Laboratory
Kenneth A. Ross, University of Oregon
Hugo Rossi, University of Utah
Shmuel Winograd, IBM

OTHER COMMITTEES

LOCAL ARRANGEMENTS COMMITTEE

Chairman: John W. Addison, Jr., University of California, Berkeley
Mary Ann Addison
Henry L. Alder, University of California, Davis
Lenore Blum, Mills College
William G. Chinn, San Francisco, California
Ginette Henkin
Irving Kaplansky, Mathematical Sciences Research Institute,
Berkeley



Hirzebruch, Alavi, Atiyah

THE CONGRESS AND ITS COMMITTEE STRUCTURE

Robion C. Kirby, University of California, Berkeley
Eugene L. Lawler, University of California, Berkeley
Calvin C. Moore, University of California, Berkeley
Robert Osserman, Stanford University
P. Emery Thomas, University of California, Berkeley
Anthony J. Tromba, University of California, Santa Cruz

COMMITTEE ON SPECIAL FUNDS

Committee Members:

Chairman: Gerald L. Alexanderson, University of Santa Clara
Richard D. Anderson, Louisiana State University
Morton L. Curtis, Rice University
Susan J. Friedlander, University of Illinois, Chicago
David Gale, University of California, Berkeley
Andrew M. Gleason, Harvard University, (ex-officio)
Daniel Gorenstein, Rutgers University, New Brunswick
Jill P. Mesirov, Thinking Machines Corporation
(ex-officio)

Executive Secretary: K. Brooks Reid, Louisiana State University
Alice T. Schafer, Wellesley College and Simmons College
Albert R. Stralka, University of California, Riverside

Advisory Board:

George E. Brown, Jr., Member of the United States
Congress, 36th District, California
Edward E. David, Jr., President, Exxon Research and
Engineering Company
Gerard Debreu, Professor of Economics and Mathe-
matics, Nobel Laureate in Economics, University of
California, Berkeley
Marvin L. Goldberger, President, California Institute
of Technology
Ira Michael Heyman, Chancellor, University of
California, Berkeley
Brockway McMillan, Vice-President, Military Systems,
Bell Laboratories, retired; former Under Secretary,
U.S. Air Force
I. M. Singer, John D. MacArthur Professor of Mathe-
matics, Massachusetts Institute of Technology
James M. Vaughn, Jr., Chairman, Vaughn Foundation
Fund

PUBLIC INFORMATION COMMITTEE

Director of Publicity: Donald J. Albers, Menlo College
Yousef Alavi, Western Michigan University

Advisor: Gerald L. Alexanderson, University of Santa Clara
William G. Chinn, San Francisco, California
Jane M. Day, San Jose State University

THE CONGRESS AND ITS COMMITTEE STRUCTURE

Advisor: Ronald L. Graham, Mathematical Sciences Research
Center, AT&T Bell Laboratories
Peter J. Hilton, SUNY at Binghamton
Lester H. Lange, San Jose State University
Jean J. Pedersen, University of Santa Clara

ADMINISTRATIVE STAFF

Executive Director: Jill P. Mesirov, Thinking Machines Corporation
Congress Manager: H. Hope Daly, American Mathematical Society
Business Manager: Lee Ann Lima

LIST OF DONORS

GRAND BENEFACTORS

Department of the Air Force, Air Force Office of Scientific Research
The Board of Trustees of the American Mathematical Society
The Membership of the American Mathematical Society
Department of the Army, Army Research Office
Department of Energy, Office of Energy Research
Department of the Navy, Office of Naval Research
National Science Foundation
Alfred P. Sloan Foundation
University of California, Berkeley
James M. Vaughn, Jr., Vaughn Foundation Fund

BENEFACTOR

International Mathematical Union

PATRONS

AT&T Foundation
Exxon Research and Engineering Company
GTE Laboratories Incorporated
General Motors Corporation
Grumman Corporation
John Hancock Mutual Life Insurance Company
International Business Machines Corporation

SPONSORS

Amoco Foundation, Inc.
Honeywell Inc.
Joint Policy Board for Mathematics
Mathematical Association of America
James H. Simons

DONORS

Deloitte Haskins & Sells
HERMANN publishers in arts and sciences
Hughes Aircraft Company
Pergamon Press
Raytheon Company
Springer-Verlag
Texaco Philanthropic Foundation Inc.
Western Michigan University

CONTRIBUTORS

American Microsystems, Inc.
Anonymous Contributor
Associated Microfilm Management Services, Inc.
Battelle Memorial Institute
Leonard E. Baum
Chevron Research Company
Chrysler Corporation
Computer Sciences Corporation
Cornnuts, Inc.
Institute for Defense Analyses, Communications Research Division
Lockheed Corporation
Louisiana State University
Mother's Cookies
The MITRE Corporation
Society for Industrial and Applied Mathematics
University of Illinois at Chicago

FRIENDS

Richard D. Anderson
Anonymous in Appreciation of Hope Daly and Staff
Applied Technology
Association for Women in Mathematics
Bowerman Brothers, Inc.
Cooley, Incorporated
Data Tech's, Inc.
DES Offset, Inc.
Harvard University, Department of Mathematics
Hasbro, Inc.
Industrial Electric Company
Lyons & Morrison Paper Co.
Massachusetts Institute of Technology, Department of Mathematics
Mathematical Reviews
Modern Press, Inc.
E.P. Miles, Jr.
Polaroid Foundation, Inc.
Precision Business Forms, Ltd.
Rourke-Eno Paper Company, Inc.
Andrew Shahinian
Simmons College
Thinking Machines Corporation
University of Santa Clara
Wellesley College
Wesleyan University

OTHERS

Kailash K. Anand

Truman A. Botts

Dreyer's Grand Ice Cream

East Bay Municipal Utility District (California)

Freeman Furniture, Inc.

Michael Integlia, Jr.

Bill Israel, Conference Marketing Coordinator,

Jolly Giant Conference Center, Humboldt State University

Murray S. Klampkin and Andrew C-F Liu, University of Alberta

Massachusetts Envelope Company

Port of Oakland (California)

RIS Paper Company, Inc.

Edward F. Schmeichel

Joseph R. Tassone

United Paper Stock Company

The City of Berkeley extends a hearty welcome to the participants and their accompanying persons to the International Congress of Mathematicians and wishes them success in their science, happiness in their personal lives, and the satisfaction of helping in the development of science for peace and human well-being.

Eugene "Gus" Newport
Mayor of Berkeley

It is a pleasure to extend warm regards on the occasion of your 1986 Congress. Mathematics plays an increasingly significant role in every aspect of our lives. This prestigious conference provides a fine opportunity for participants to share ideas regarding important services which involve mathematical concepts and theory. The impressive array of speakers and participants promises to provide a comprehensive examination of mathematics in today's world. As you gather to exchange information, may reflection on past achievements provide strength to meet the challenges of the future. Please accept my best wishes for an informative conference. Should you find time between sessions to explore the beautiful San Francisco Bay area, I am sure you will find the pleasant hospitality of this great state awaiting.

Most cordially,
George Deukmejian.
Governor of California

OPENING CEREMONIES

The opening session of the Congress was held in the Greek Amphitheatre of the University of California, Berkeley, at 9 a.m. on August 3, 1986.

The New Albion Brass Quintet played "Mini Overture for Brass Quintet" by Witold Lutoslawski.

Professor Jürgen Moser, President of the International Mathematical Union, opened the Congress with the following words:

On behalf of the International Mathematical Union I wish to welcome you to the International Congress of Mathematicians 1986. It is one of the primary functions of the IMU to support and promote the International Congress that is held every four years.

It may be good to recall that these Congresses have a long history going back to the last century. The first one was held in Zurich in 1897, the second in Paris in 1900. However, the sequence of Congresses was broken by the two World Wars. Now, since the Congress of 1950 in Cambridge, Massachusetts, at which the IMU was founded, we have been in the fortunate position to have had a long, uninterrupted sequence of Congresses. Today we are here to begin the tenth Congress in this series, and I am sure we are all united in the fervent hope that this series will continue uninterrupted well into the next century.

At the first Congress there were 216 participants; today more than 3000 mathematicians are attending this Congress. But regardless of this impressive increase, the Congress is still guided by the same principle: to foster personal relationships between mathematicians from different countries and to present a survey of the present state of science.

At a time of increasing specialization and of rapid proliferation of mathematics into many subfields, these Congresses play a particularly important role in bringing together mathematicians of different interests and backgrounds. The danger of fragmentation of our science into many separate branches cannot be overemphasized. It is our hope that this Congress will help to counter this divisive tendency and give us a wide perspective of mathematics.

I am happy to greet mathematicians from about seventy countries here in Berkeley. I hope that the coming week will provide many occasions for fruitful exchanges of ideas as well as for lasting scientific and personal contacts.

The organization of the Congress 1986 lies in the competent hands of the Organizing Committee associated with the American Mathematical Society. The Chairman of its Steering Committee is Professor Andrew Gleason. I propose that we elect, here and now, by acclamation, Professor Gleason as President of the Berkeley Congress 1986.



Opening Ceremonies, Gleason, President of the Congress

Professor Gleason was elected and spoke as follows:

It is truly an honor to preside over this, the twentieth International Congress of Mathematicians. It is also a great pleasure to welcome you all to the City of Berkeley, a city made famous by the University of California and a city where the weather is almost always as pleasant as it is this morning. Speaking on behalf of the National Academy of Sciences and on behalf of the entire mathematical community I extend a special welcome to all those who have come from other countries. May your trip to the United States be a pleasant one, may you learn some new mathematics, make some new friends, and enjoy some of the marvelous sights in this vast country.

As you know, the Congress has been organized into nineteen sections. The Program Committee, under the able chairmanship of Professor Fritz Hirzebruch and with the aid of numerous subcommittees, has prepared an extraordinary scientific program with 16 plenary speakers and over 140 sectional speakers. In addition there are more than 400 contributed papers, and several specialized seminars have already been set up. We regret that a few of the invited speakers have been unable to come and that we have had to make some last-minute changes in the schedule. We will keep you informed of all the changes through the daily newsletter.

Mathematics has always been useful. Many of the oldest written records of human civilization are accounting documents, and in fact today accounting still is the largest application of mathematics. But we are rapidly moving into a period in which more and more applications of mathematics are being found. New mathematical questions are being asked by scientists, engineers, and managers—often questions of an entirely different sort from those previously considered. New mathematical answers are being found often involving ideas previously thought to be entirely abstract and utterly nonutilitarian. As mathematicians we can justly be proud that the concepts we have worked so hard to develop are helping people to understand the real world just as they have helped us to understand our platonic world. There is a lesson in this, I think, and it is that, as we enter this new era dominated by computers, we should not fall into the trap of utilitarianism, but remember that the greatest progress in mathematics is always made by trying to understand the fundamental structures that underlie the subject rather than attempting to solve purely utilitarian problems.

He then recognized Professor Mary Ellen Rudin, Chairman of the United States National Committee for Mathematics, who said:

I would like to welcome all of you for the United States also. Exactly fifty years ago at the Congress held in Oslo, Norway, the first Fields Medals were awarded. The two 1936 Medals went to Jesse Douglas, who is no longer living, and to Lars Ahlfors, who was then a young man not yet thirty years of age. In special celebration of the fiftieth anniversary of the Fields Medal and of Professor Ahlfors' fifty years of continued contributions to mathematics, I would like to nominate Professor Lars Ahlfors to be Honorary President of the Congress.



**Opening Ceremonies: Faddeev, Mesirov, Moore,
Johnson, Rudin, Ahlfors, Moser, Gleason**

Professor Ahlfors was elected by acclaim and came to the lectern and said:

I accept this great honor with a good conscience because I consider myself a link between this International Congress and the one in 1936, fifty years ago, the occasion on which the Fields Medals were given for the first time. I understand that my only duty here will be the pleasure of handing out the Fields Medals and the Nevanlinna Prize.

At that time the circumstances were quite different; the idea of the medals had been approved in Zurich in 1932, but there had been no publicity about it and when I arrived in Oslo I did not know that the Medal had become a reality, and if I had known it I would not have considered myself the right candidate. As a matter of fact, I had not been told anything officially until I entered the room where the opening ceremony would take place, but there I was shown a place somewhere in front, and I may have had my suspicions. Well, I had more than that. I had been warned beforehand by somebody who by mistake congratulated me a day before. But up to that point it had been a secret at least officially, even to myself. There was no tradition to go by and no protocol to follow. As was mentioned here, two medals were given, one to me and one to Jesse Douglas, who was then at MIT while I happened to be a visiting lecturer at Harvard. In that way it so happened that both medals went to Cambridge, Massachusetts. Unfortunately Douglas could not accept his medal in person because according to the Congress record he was too tired. I don't know, but maybe he had good reason to be tired after a long and strenuous journey. I would not expect that to happen today. His medal was then accepted by Norbert Wiener as representative of MIT.

There are two traditions that go back to the very beginning. In the first place, the Committee to select the winners should consist of the top brass of contemporary mathematics. In 1936 the members of that Committee were G.D. Birkhoff, Caratheodory, Elie Cartan, Severi, and Takagi. Truly I would call that a panel of Olympian heroes. And I think that this tradition has been continued at subsequent Congresses. The other tradition is that the works of the winners should be commented on by prominent persons in the field. In 1936 both prizes were explained by Caratheodory.

As was mentioned there was no Congress until 1950, fourteen years later. On that occasion, which took place at Harvard, the medals were given to Atle Selberg and Laurent Schwartz, both known and admired by all mathematicians. From then on the Fields Medals have become more and more prestigious and it is a safe bet that many dream of getting it. Whether true or not that the existence of the medal has contributed to the phenomenal growth of mathematics both in quantity and quality during the last fifty years must remain anybody's guess.

Today it is safe to congratulate the winners in advance and I use this occasion to offer them my sincerest compliments to their success. I share their feeling of pride and accomplishment and I know that their continued success is guaranteed. I also share the disappointment of the many who may feel that they have been passed by. I wish them better luck next time or, if there is not a next time, that posterity will prove them right and the Committee wrong. Thank you.

The brass quintet then played selections from "A Brass Menagerie" by John Cheetham.

Professor Gleason then introduced Professor Calvin C. Moore, Associate Vice President for Academic Affairs of the University of California, who spoke as follows:

The President of the University of California, David Gardner, asked me to convey his regrets that he could not be here in person and to tell you how pleased he is that the University of California is hosting this distinguished International Congress. I am delighted to represent him on this occasion for at least two reasons. First of all, I am a mathematician myself, so it is a special pleasure to help with the official ceremony of a week of addresses, lectures, and opportunities for discussion with the most distinguished mathematicians in the world. As President Gardner noted in his message in the program of this opening ceremony, international cooperation has expanded in most disciplines, but nowhere has such cooperation flourished more than in the field of mathematics. Indeed, mathematics today is brimming with new ideas and developments; so this Congress promises more than its share of stimulation and excitement. Second, in my capacity as Associate Vice President for Academic Affairs of the University of California, I am pleased to welcome you to the University of California. I am also a faculty member here on the Berkeley campus and I can assure you that, even though the founders of the City of Berkeley named it after a philosopher rather than a mathematician, Berkeley has been a hospitable environment for mathematicians. The University is proud of its highly regarded Department of Mathematics and of the newly created Mathematical Sciences Research Institute, and I should add that Berkeley is situated in one of the most interesting and lively areas in the United States. I hope very much that you will take the time to explore some of the cultural and other highpoints of the San Francisco Bay area during your stay. I think you will find it a stimulating and fascinating place. But my duty is not just to welcome you to the Berkeley campus, but to the entire University of California. Therefore, on behalf of the University and its Board of Regents, its Presidents, its Chancellors, its 9 campuses, and 150 Centers, Institutes, Laboratories, and Bureaus throughout California, its 148,000 students and 9,000 faculty members, including an aggregate of mathematics faculty of 450, I bring you greetings and warmest good wishes for a pleasant visit and a most successful conference.

Professor Gleason then read from a proclamation by Eugene Gus Newport, Mayor of the City of Berkeley, and a message from George Deukmejian, Governor of the State of California. He then introduced Richard Johnson, Acting Science Advisor to the President who spoke as follows on behalf of President Ronald Reagan:

On behalf of the American people, I am pleased to welcome you to the United States. For only the second time in the twentieth century—the century that history will remember best for the remarkable discoveries and tremendous advances made in science and technology—this country

is privileged to host the best mathematicians from some seventy nations around the world. I would like to use this occasion to congratulate, in advance, the winners of the Fields Medals and the Nevanlinna Prize. You recipients know who you are, but the suspense is beginning to wear on the rest of us, who are eager for the announcement and the opportunity to give you the accolades and recognition you so richly deserve.

Here in the United States, and I'm sure in your countries as well, my colleagues in science and technology engage in a certain friendly rivalry between our various disciplines. For example, I might say: "Physics is the most important of the sciences because it seeks to uncover the secrets of the tiniest particles of matter which make up our universe." The biologist might reply, "What could be more significant to humanity than understanding the building blocks of life itself." Or the chemist: "Don't underestimate the value of chemistry." But we scientists are not a contentious bunch by nature, and I can think of at least two points on which American scientists would echo resounding accord: First, science and technology are essential to improving the quality of life for all mankind, and second, mathematics is essential to every field of science and technology.

Of course, for me, it is ample honor to stand before this august gathering as a representative of the United States government, but I derive great personal pride and pleasure from the opportunity this event provides me, as a representative of the American scientific community, to express the gratitude of my fellow scientists to the mathematicians of the world, whose tireless efforts and manifold contributions form the very foundation upon which our work is built. We owe a large measure of our success to you. As science and technology progress at a breakneck pace—with mathematics as the fundamental driving force—we will continue to look to you for guidance, direction, and assistance. But today, I would also like to urge you to take on a role beyond solely enhancing your own research capabilities and contributions.

Throughout the world, science and technology increasingly are acknowledged to be essential to a nation's economic strength, industrial productivity, national security, and overall quality of life. While science is never static, right now we are in the midst of an unprecedented era of revolution in virtually every field of science and technology—a revolution that we created, and one which we have a responsibility to sustain and, indeed, to lead. The most important point, and the principal source of my concern, is that we must also possess the ability to respond to the opportunities afforded by this rapidly progressing scientific environment. To do so will require a strong base of talent—men and women, like you, who will lead the world's science and technology talent enterprise into the unforeseeable reaches of the twenty-first century. The pressing question is whether or not we will have the science and technology talent base adequate to meet the challenges of the future. Looking at current statistical indicators, the alarming answer, in most countries, is no.

We in the United States recognize that we have a serious problem. Fewer and fewer young people are studying science, mathematics, engineering, and technology; fewer and fewer young people are pursuing advanced de-

grees in these disciplines; and fewer and fewer young people are choosing careers in the scientific and technological fields. If this trend is allowed to continue, this country, and others similarly afflicted, will be unequipped to maintain the level of scientific leadership on which our improving standard of living, and the more general quality of life, depends.

To my mind, the most important aspect of this Congress is its international nature. Like mathematics itself, it transcends the barriers of language and ideology, the boundaries of geography, and the adversarial climate of competition. You are convened here this week, in the spirit of cooperation, to share your latest discoveries and to learn about new developments, and to pursue your mutual interests and the solutions to common problems. Perhaps that is what compels me to share this growing problem of an inadequate talent base with you, and to turn to you for help in solving it.

All of us who are not mathematicians feel a keen sense of envy at the abilities you possess. The intrinsic beauty (especially as now seen in 3-D color), the coherence, the elegance of mathematics makes it an entirely worthwhile pursuit in its own right. You have the rare talent to appreciate fully the depth and power of your subject. But, with ability and talent come obligations and responsibilities, especially to those not so blessed. What you do in your laboratories, at your computer terminals, and in your studies is remarkable—indeed, it has changed the course of the world. But you must do more. It is no longer enough to be practitioners of science. You must also become citizens of science—leaders of a uniquely gifted community, bound by the common goal of assuring that the progress you have made and the improvements you have wrought for humanity will continue. Especially, you must help to assure that there is an expanding, well-qualified generation of scientists, engineers, and technologists to succeed you.

What the young students of the world need most, in my view, is someone to imbue them with the excitement and enthusiasm that brought you here today, someone to communicate to them the importance of science and technology to society, to demonstrate to them that mathematics, physics, biology, chemistry, and their sister disciplines are challenging pursuits, worthy of their highest interest and attention. Many young teachers, I hasten to add, need to hear this as well. No one is better qualified to carry those messages than the individuals who have made eminently successful careers out of the quest for knowledge. In short, the best spokesmen for mathematics, physics, biology, and chemistry are mathematicians, physicists, biologists, and chemists.

Unfortunately, too many students, especially at the grammar school level, still regard science and mathematics as “too hard,” too difficult. To them, these subjects look ponderous, abstract, and impenetrable, and many children dismiss them before they’ve had a chance to appreciate the fun and excitement inherent in the challenge of discovery. At that moment, a potential scientist or mathematician is lost—for now, more than ever before, the attitudes and academic foundations developed in the early school years determine the future course of a student’s education and career. In the current scientific and technological climate, with a growing demand

for an expanded talent base, we cannot afford to lose these students in grammar school—well before there is a need for a carefully weighed career choice.

Therefore I believe it is incumbent upon you to become active participants in the community of science, to help others to grasp the broad importance of science and technology to the very fabric of our daily lives—to help them understand the momentous potential of science and technology to enhance the quality of life for all mankind. Certainly you can point with pride to the eradication of diseases, to the agricultural revolution, to hurricane forecasts, to the transportation, automation, electronics, and information revolutions, as examples. We must, of course, recognize and limit the risks associated with modern uses of science and technology. But, further, we must convey to our young students not only factual information on the present extensive benefits, but also the great challenges, personal opportunities, and potential future benefits associated with the scientific and technological disciplines.

I urge you to get involved in developing the talent bases needed in the twenty-first century for expanding the contributions of science and technology to humanity. On a small scale, you can, of course, accomplish a great deal simply by going to your local schools and talking to the teachers and the children about the contributions that they have the power to make through science. On a larger scale, you can work through existing mechanisms, like your professional societies, and enlist the help of your colleagues in devising innovative solutions to this set of perplexing problems. That, after all, has been your mission within the discipline of mathematics; now I challenge you with this added responsibility to mankind.

Individuals of the highest caliber must be sought out, trained, nurtured, and encouraged by every means possible to carry on the tradition of excellence you represent. I can think of none better suited to the task than you, who, by your example, your actions, and your words can express the merit and rewards of a life in science.

In closing, I have a personal message from a man who is himself excited by science and technology and firm in his commitment to their invaluable role in international health and prosperity. [The text of President Reagan's message may be found on page XXXIII.]

Professor Gleason thanked Dr. Johnson and asked him to convey the thanks of the meeting to President Reagan. Then the brass quintet played "Fanfare" by David Amram. Professor Gleason announced that Professor Ahlfors would present the medals, and then called on Academician Ludwig Faddeev, the Chairman of the Committee to Award the Nevanlinna Prize, who responded as follows:

The Nevanlinna Prize has a much shorter history than the Fields Medal, as we have heard from Professor Ahlfors. It was established to a great extent due to the efforts of our past president, Lennart Carleson, with the aim of stressing the importance of the applied aspects of mathematics. Helsinki University generously provided necessary funds, so it is only appropriate that it is called the Nevanlinna Prize. More exactly, the prize is to be awarded to a young mathematician for outstanding work in the mathe-

mathematical aspects of information science. "Young" means the same as for Fields Medals. The Prize is awarded for the second time. The first award was given in the Warsaw General Assembly. The Committee for the award for this year consisted of Professor Cook, University of Toronto; Professor Winograd, IBM; and myself. Let me inform you of our decision. The 1986 Nevanlinna Prize is awarded to Leslie Valiant, Harvard University, USA. I will ask Professor Ahlfors to present the Prize.

Professor Valiant then came forward to receive the Nevanlinna Prize from Professor Ahlfors.

Professor Gleason then recognized Professor Moser in his capacity as Chairman of the Committee to Award the Fields Medals and Professor Moser reported as follows:

In the early 1930s the organizers of the Toronto Congress decided to award at each Congress two gold medals for outstanding achievements in mathematics. The driving force behind this resolution was John Charles Fields who expressed the wish that this prize be given, I quote, "in recognition of work already done and also as an encouragement for further achievements on the part of the recipients and as a stimulus to renewed efforts on the part of others." It is interesting to note that Fields insisted "that the



Musicians at the Opening Ceremonies

medals be of a character as purely international and impersonal as possible." As a matter of fact, Fields' name does not appear on the medals. As we heard today already, the first medals were awarded at the 1936 Congress in Oslo, four years after Fields' death. Looking at the impressive list of Fields Medalists over the past fifty years, it is obvious that Fields' vision has been realized most successfully.

Since the Congress in Stockholm in 1962, the International Mathematical Union has been entrusted with the selection of the Fields Medal winners. The Committee for the Fields Medalists for this Congress consisted of P. Deligne, J. Glimm, L. Hörmander, K. Ito, J. Milnor, S. Novikov, C. S. Seshadri, and myself as Chairman. The Committee had the difficult task of making a selection from an impressive list of brilliant mathematicians. It goes without saying that such a decision has no unique solution and that all the excellent mathematicians considered could be awarded the prize. On behalf of the International Mathematical Union, I want to thank the members of the Committee for their efforts to reach the decision in a spirit of responsibility and cooperation.

The Committee followed the tradition of limiting the awards to mathematicians under forty years of age. After long deliberation and extensive consultation it was decided that three young mathematicians be selected for this award. We are all deeply impressed by their exceptional achievements and I am happy to report that the Committee was unanimous in its support of the three Fields Medalists for 1986. Their names are Simon Donaldson, Gerd Faltings, Michael Freedman. I offer them our warmest congratulations. I now ask the winners to come up to the podium.

The winners came forward and received their medals from Professor Ahlfors, and then the brass quintet played selections from "Three Pieces for Brass Quintet" by Minoru Fujishiro.

Professor Gleason congratulated the winners and announced that talks concerning their work would be given in Zellerbach Auditorium.

The session adjourned at 11:30 a.m.

THE WHITE HOUSE
WASHINGTON

July 3, 1986

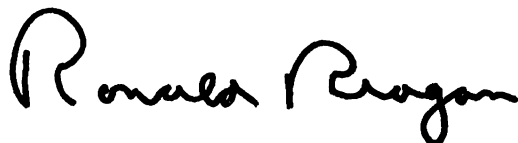
I extend a warm welcome to the thousands of mathematicians from around the globe who are attending the quadrennial International Congress of Mathematicians. All of us are pleased and, indeed, honored that the United States was chosen to host this prestigious meeting.

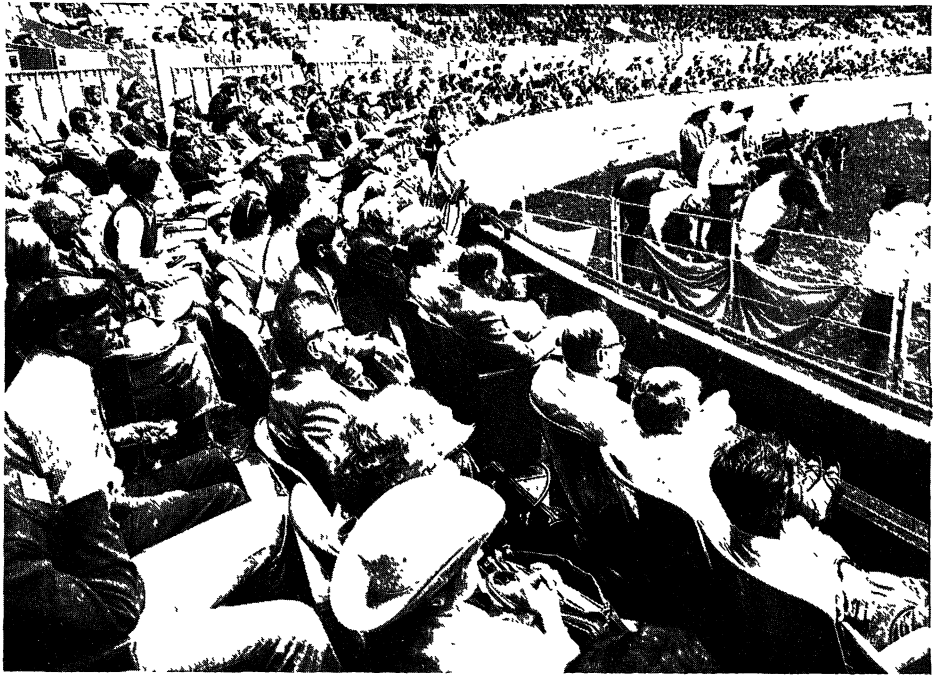
Mathematics is the enabling force for the revolutionary advances being made throughout the world in science and technology. The fundamental role of mathematics is becoming increasingly apparent in business, industry, and government. Modern mathematicians are giving new meaning to the famous tenet of the ancient Pythagoreans that "all is number."

I am gratified to note that this Congress will be the occasion for the awarding of Fields Medals to outstanding members in recognition of their contributions to mathematics. I wish to extend my congratulations to the winners.

It is appropriate that these honors be presented at an international meeting, for mathematics is intrinsically international, cutting across geographical and cultural boundaries with its own language. International competition is a concept alien to the study of mathematics. Indeed, cooperation between mathematicians of different nations has been a long-standing tradition.

I wish you the best for a successful meeting. God bless you.

A handwritten signature in black ink, reading "Ronald Reagan". The signature is written in a cursive, flowing style with a large, prominent "R" at the beginning.



The Rodeo

CLOSING CEREMONIES

The final session of the Congress was held in Zellerbach Auditorium at 11 a.m. on August 9.

Professor Gleason opened the meeting and called upon Professor Moser for a report on the International Mathematical Union. Professor Moser spoke as follows:

We have come to the end of this Congress with a rich and interesting program. At this occasion it is a tradition for the President of the International Mathematical Union to address the Congress and to report on the work and the decisions taken at the General Assembly.

Before doing so, it is my privilege to greet here Professor Marshall Stone, one of the past presidents of the International Mathematical Union. It was Professor Stone who played a decisive role in reestablishing the IMU in 1950 after it ceased to exist in 1932. Professor Stone, we are happy to have you with us at this Congress.

Since 1962 the IMU has been responsible for preparing the scientific program of the Congresses. This task lies in the hands of the Program Committee, which is appointed in part by the IMU and in part by the host country. For the present Congress this Program Committee consists of Fritz Hirzebruch (chairman), Enrico Bombieri, Lennart Carleson, David Mumford, Louis Nirenberg, Michael Rabin, Yuri Rozanov, David Ruelle, and Isadore Singer. I would like to express our thanks to Professor Hirzebruch and his Committee for presenting us with an excellent program of great diversity.

It was a great disappointment for all of us, however, that many of the invited speakers from the Soviet Union did not come to Berkeley; in fact, almost half of the Soviet speakers were not present. This is a serious loss for everybody concerned and defeats the purpose of the Congress. It is most important for any Congress that the invited speakers are able to attend in order to deliver their lecture in person and to take part in the exchange of ideas.

We are aware that our Soviet colleagues worked very hard at resolving this problem, and we appreciate their efforts. Also, most of the manuscripts of the absent speakers were made available and could be presented by other mathematicians.

Regardless of circumstances, it is always a disappointment if invited speakers from any country are unable to attend, and let me express our hope that at the Congress 1990 all invited speakers from all countries will be present.

As you may know, the IMU is a member of the International Council of Scientific Unions (ICSU) and as such is committed to the ICSU principle of free circulation of scientists. I am happy to report that to the best of my



Moser and Faddeev: outgoing and incoming presidents of IMU



Chancellor's Reception

knowledge the host country has granted all visas which have been applied for. In some difficult cases the help from ICSU was indeed essential. This again demonstrates the importance of the ICSU principle for our Union. Let me add that two weeks ago at the General Assembly of the IMU a resolution was adopted reaffirming an ICSU article on nondiscrimination.

I will now give a brief report on the General Assembly of the IMU. On July 31 and August 1, the General Assembly met in Oakland, California, and I want to inform you about the main decisions taken there.

Two new members were accepted by the General Assembly, and Ivory Coast as well as the People's Republic of China belong now to the IMU. Let me mention that the question of the China membership has a long and complicated history. Here I want to thank especially the Secretary of the IMU, Professor Olli Lehto, for his untiring efforts to bring this problem to a successful solution. We also want to express our appreciation to China-Taiwan for its cooperation in helping to solve this problem.

The total number of IMU members is now fifty-three. Furthermore, the General Assembly elected new Committees and I want to report to you the outcome. I begin with the new IMU Executive Committee for 1987-1990. It consists of the President, Ludwig Faddeev; the Vice-Presidents, Walter Feit and Lars Hörmander; the Secretary, Olli Lehto; and the members, John Coates, Hikosaburo Komatsu, László Lovász, Jacob Palis, Jr., and C. S. Seshadri. To this is to be added the Past President as ex officio member.

Now I come to the subcommissions. The IMU has two subcommissions: the International Commission on Mathematical Instruction (ICMI) and the Commission on Development and Exchange (CDE). The new Executive Committee of ICMI is as follows: the President, Jean Pierre Kahane; two Vice-Presidents, Peng-Yee Lee and Emilio Luis Riera; the Secretary, A. G. Howson; and the three members are Hiroshi Fujita, Jeremy Kilpatrick, and Mogens Niss. The new CDE consists of the Chairman, M. S. Narasimhan, and the members Jean Pierre Bourguignon, Phillip Griffiths, M. Immanaliev, A. O. Kuku, Lê Dũng Tráng, Shingo Murakami, and Giovanni Vidossich. To the ICMI Executive Committee and the CDE, the IMU President and Secretary have to be added as ex officio members.

Finally, the Site Committee made its proposal to the General Assembly about the location and time of the next Congress.

Before giving the word to the next speaker, let me conclude with some personal remarks. In my work during the past four years as President, I was fortunate to have the advice and support from many colleagues throughout the world. I would like to thank all of them—in particular, the members of the Executive Committee and the Fields Medals Committee. The strongest support, however, came from Olli Lehto, who together with his secretary, Mrs. Tuulikki Mäkeläinen, was always ready to help at any time in every difficulty. Without his encouragement and sensitive judgment I would not have carried out this task. To both of them go my warmest thanks.

Professor Moser then yielded the floor to Professor Nagata who invited the audience to the next International Congress with these words:

On behalf of the Japanese Committee for Mathematics, I have the honor



Nagata inviting the Congress to Kyoto

of inviting you to the next International Congress of Mathematicians in Kyoto. Kyoto had been the capital of Japan for about one thousand years and can show you some of the old Japanese culture.

We are quite aware that it must be a difficult task to organize such a big meeting. However, taking into account the help of the International Mathematical Union and also the cooperation of the mathematical community of the world, I believe that we will be able to overcome the difficulties.

Hoping that you will accept our invitation, I welcome all of you to the next International Congress of Mathematicians to be held in August 1990 in Kyoto.

Professor Gleason then recognized Professor Adams who said:

Ladies and gentlemen, on such occasions it is right and proper for guests to thank their hosts.

We, the mathematicians of the world, have come to Berkeley from our many countries. We have had a most successful conference, and we wish to express our appreciation. We have had the chance to meet and speak to each other; we have heard many splendid lectures and thanked the speakers with our applause. To arrange all this has taken many people much work.

As Professor Moser has explained, to choose the speakers was the task of a committee structure set up by the IMU, and presided over by Professor Hirzebruch. We know that mathematicians of all countries give their services in this most gladly; we still thank them again for making an informed choice on our behalf.

Almost all the work, however, falls upon the hosts. For this purpose the American Mathematical Society set up a committee structure presided over by Professor Gleason, and we thank all those involved. We also thank the permanent staff of the American Mathematical Society, whom you may have seen at the Congress Bureau. These people have had to book accommodation, arrange all the things that need arranging, and assume the responsibility of running a fair-sized business. (By the way, there are still a few cowboy hats left for sale) The praise and gratitude due to them will best be estimated by those who have tried anything of the sort before, for example Professors Czeslaw Olech, Olli Lehto, and Maurice Sion. If the rest of us would regard smaller exercises of the kind with apprehension, our applause will start on the right foot.

I ask all the mathematicians of other countries to join me in thanking our American hosts, and all of us to thank those who have done the work.

After thanking Professor Adams, Professor Gleason spoke as follows:

There are two principal ingredients in a successful Congress, the scientific program and the physical arrangements. The Program Committee chose an excellent roster of speakers, a very distinguished roster indeed, and I thank them for doing so. And I thank the speakers, many of whom came from afar to share their insights with us. The lectures were exceptionally fine. The program was further enriched by several special seminars and many contributed papers. The organizers and speakers deserve our thanks.



Hope Daly, Congress Manager

The success of a scientific meeting depends on the participants. Over 3500 mathematicians attended this Congress. The total registration was 3970. Thank you all for coming. Speaking for the American mathematical community, I want to thank particularly those who came from other countries. We were honored by your presence and we wish you great success in solving the problems that we have heard about in the lectures.

There are many members of the American mathematical community who worked hard for this Congress during the past two years—the list is so long that I cannot mention them all—but I would like to express my personal thanks and also to relay the generous words of Professor Adams, which I believe truly represent the feelings of the membership. I do want to mention the Steering Committee and, in particular, Kenneth Ross and Hugo Rossi, who were in charge of scheduling; the Board of Directors, under the chairmanship of Ronald Graham; the Committee on Special Funds, chaired by Richard Anderson; the Public Information Committee, chaired by Yousef Alavi; and the Local Arrangements Committee, chaired by John Addison. All of these Committees have pulled more than their weight and the results are manifest.

Among the mathematicians who worked for this Congress is one who deserves very special mention, Jill Mesirov, the Executive Director of ICM-86, who has been in daily contact with the preparations for the Congress since the beginning; we owe her a special vote of thanks.

I want to acknowledge here the outstanding contribution of the organizers of the Warsaw Congress. They received our Congress Manager most cordially in 1983 and shared their experiences with her. Their advice was invaluable.

Finally, I want to thank the Congress Manager, Hope Daly. She has been our field marshal for the past four years, directing all of the preparations, yet never shrinking from the minor tasks when an extra hand was needed.

Professor Gleason then presented Ms. Daly with a silver pendant and a bouquet. There was much applause, after which he continued:

This Congress is part of a long tradition of internationalism. At least since the days of Archimedes, mathematicians have corresponded with one another and traveled great distances to study, teach, and confer. As the expense of printing and traveling has declined, the tradition has strengthened. Now hundreds of mathematical books and journals are published every year. These pass freely over international boundaries and propagate new mathematical ideas throughout the world. Mathematicians travel ever more frequently from one university or institute to another. As we think of this Congress, let us resolve to maintain and expand our great tradition of freedom to study, travel, and confer so that the Kyoto Congress will be even more truly international.

Professor Gleason then declared the Congress closed.



Zellerbach Auditorium

SCIENTIFIC PROGRAM

PLENARY ADDRESSES BY INVITATION OF THE ORGANIZING COMMITTEE

DE BRANGES, L., Underlying concepts in the proof of the Bieberbach conjecture.....	25
DONALDSON, S. K., The geometry of 4-manifolds.	43
FALTINGS, G., Neue Entwicklungen in der arithmetischen algebraischen Geometrie.....	55
FRÖHLICH, JÜRG (Paper not read at the Congress. No manuscript received.), Analytical approaches to quantum field theory and statistical mechanics.....	62
GEHRING, F. W., Topics in quasiconformal mappings.....	81
GROMOV, M., Soft and hard symplectic geometry.	99
LENSTRA, H. W., Elliptic curves and number-theoretic algorithms.....	121
SCHOEN, R., Recent progress in geometric partial differential equations.	131
SCHÖNHAGE, A., Equation solving in terms of computational complexity.	154
SHELAK, S., Taxonomy of universal and other classes.	163
СКОРОХОД, А. В., Случайные процессы в бесконечномерных пространствах. (Random processes in infinite dimensional spaces.)	172
SMALE, S., Algorithms for solving equations.	196
STEIN, E. M., Problems in harmonic analysis related to curvature and oscillatory integrals.	222
SUSLIN, A. A. (Paper read by E. Friedlander), Algebraic K -theory of fields.	245
VOGAN, D. A., Representations of reductive Lie groups.....	267
WITTEN, E., Physics and geometry.....	

ADDRESSES IN SECTIONS BY INVITATION OF THE ORGANIZING COMMITTEE

The Sections

01 Mathematical Logic and Foundations	307
02 Algebra.....	338
03 Number Theory.....	417
04 Geometry	484
05 Topology	574
06 Algebraic Geometry	623
07 Complex Analysis	699
08 Lie Groups and Representations	783
09 Real and Functional Analysis	871
10 Probability and Mathematical Statistics	994
11 Partial Differential Equations and Dynamical Systems.....	1057
12 Ordinary Differential Equations and Dynamical Systems	1150
13 Mathematical Physics	1259

14 Numerical Methods and Computing	1319
15 Discrete Mathematics and Combinatorics	1400
16 Mathematical Aspects of Computer Science	1444
17 Applications of Mathematics to Nonphysical Sciences	1496
18 History of Mathematics	1612
19 Teaching of Mathematics	1668

Section 01

KECHRIS, A. S., The complexity of antidifferentiation, Denjoy totalization, and hyperarithmetical reals.	307
LACHLAN, A. H., Homogeneous structures.	314
MAGIDOR, M., Large cardinals and small sets: A survey. (No manuscript received.)	
ПЕРЕТЯТКИН, М. Г., Конечно аксиоматизируемые теории. (Finitely axiomatizable theories.)	322
WILKIE, A. J., On schemes axiomatizing arithmetic.	331
WOODIN, W. H., The two faces of infinity. (No manuscript received.) ..	

There were 18 Short Communications in this section.

Section 02

AUSLANDER, M., The what, where, and why of almost split sequences. .	338
БЕЛЫЙ, Г. В., О коммутанте абсолютной группы галуа. (On the commutator of the absolute Galois group.)	346
BORHO, W., Nilpotent orbits, primitive ideals, and characteristic classes (A survey).	350
BROUÉ, M., Théorie locale des blocs d'un groupe fini.	360
DE CONCINI, C., Equivariant embeddings of homogeneous spaces	369
GABRIEL, P., Darstellungen endlichdimensionaler Algebren.	378
MERKURJEV, A. S., Milnor K -theory and Galois cohomology.	389
POPOV, V. L. (Paper read by C. Procesi), Modern developments in invariant theory.	394
WILSON, R. L., Simple Lie algebras over fields of prime characteristic. .	407

There were 72 Short Communications in this section.

Section 03

GOLDFELD, D., Kloosterman zeta functions for $GL(n, \mathbf{Z})$	417
GROSS, B. H., Heights and L -series.....	425
HIDA, H., On p -adic Hecke algebras for GL_2	434
IWANIEC, H., Spectral theory of automorphic functions and recent developments in analytic number theory.	444
KOCH, H. V., Unimodular lattices and self-dual codes.	457
ODLYZKO, A. M., New analytic algorithms in number theory.....	466
WÜSTHOLZ, G., Algebraic groups, Hodge theory, and transcendence.....	476

There were 39 Short Communications in this section.

Section 04

BALLMANN, W., Manifolds of nonpositive sectional curvature and manifolds without conjugate points.....	484
BISMUT, J.-M., Index theorem and the heat equation.	491
BRYANT, R. L., A survey of Riemannian metrics with special holonomy groups.....	505
CHEEGER, J., On the formulas of Atiyah-Patodi-Singer and Witten.	515
COLIN DE VERDIÈRE, Y., Spectres de variétés riemanniennes et spectres de graphes.	522
ELIASHBERG, YA. M. (Paper read by J. Mather), Combinatorial methods in symplectic geometry.	531
HARDT, R. M., Singularities in some geometric variational problems....	540
MEEKS, W. H., Recent progress on the geometry of surfaces in \mathbf{R}^3 and on the use of computer graphics as a research tool.....	551
RUH, E. A., Almost Lie groups.	561
WENTE, H. C., Immersed tori of constant mean curvature in R^3	565

There were 45 Short Communications in this section.

Section 05

CARLSSON, G., Segal's Burnside ring conjecture and related problems in topology.....	574
CASSON, A. J., A survey of recent developments in three-dimensional topology. (No manuscript received.)	
MILLER, H., The Sullivan conjecture and homotopical representation theory.	580
MORGAN, J. W., Trees and hyperbolic geometry.	590
QUINN, F., Applications of topology with control.	598
SHALEN, P. B., Representations of 3-manifold groups and applications in topology.....	607
TOM DIECK, T., Geometric representation theory of compact Lie groups.	615

There were 54 Short Communications in this section.

Section 06

ARBARELLO, E., Periods of Abelian integrals, theta functions, and differential equations of KdV type.	623
BEAUVILLE, A., L'approche géométrique du problème de Schottky.....	628
CLEMENS, H., Curves on higher-dimensional complex projective manifolds.....	634
COLLIOT-THÉLÈNE, J.-L., Arithmétique des variétés rationnelles et problèmes birationnels.	641
NIKULIN, V. V. (Paper read by I. Dolgachev), Discrete reflection groups in Lobachevsky spaces and algebraic surfaces.	654
SHOKUROV, V. V. (Paper read by E. Viehweg), Numerical geometry of algebraic varieties.	672

VIEHWEG, E., Vanishing theorems and positivity in algebraic fibre spaces.	682
ZAGIER, D., L -series and the Green's functions of modular curves.	689

There were 15 Short Communications in this section.

Section 07

ALEKSANDROV, A. B., Inner functions: Results, methods, problems. ...	699
CATLIN, D., Regularity of solutions of the $\bar{\partial}$ -Neumann problem.	708
CHANG, S.-Y. A., Extremal functions in a sharp form of Sobolev inequality.	715
DOUADY, A., Chirurgie sur les applications holomorphes.	724
ГОНЧАР, А. А., Рациональные аппроксимации аналитических функций. (Rational approximation of analytic functions.)	739
KRUZHILIN, N. G., Description of the local automorphism groups of real hypersurfaces.	749
LEMPERT, L., Complex geometry in convex domains.	759
MAKAROV, N. G., Metric properties of harmonic measure.	766
WOLPERT, S. A., The geometry of deformations of a Riemann surface. .	777

There were 44 Short Communications in this section.

Section 08

BOROVoi, M. V. (Paper not read at the Congress.), Conjugation of Shimura varieties.	783
CLOZEL, L., Base change for $GL(n)$	791
DRINFEL'D, V. G. (Paper read by P. Cartier.), Quantum groups.	798
FRENKEL, I. B., Beyond affine Lie algebras.	821
GINZBURG, V. (Paper read by D. Vogan.), Geometrical aspects of representation theory.	840
KAZHDAN, D., Algebraic geometry and representation theory.	849
KUTZKO, P. C., On the supercuspidal representations of GL_N and other p -adic groups.	853
MIWA, T., Integrable lattice models and branching coefficients.	862

There were 16 Short Communications in this section.

Section 09

BOURGAIN, J., Geometry of Banach spaces and harmonic analysis.	871
CONNES, A., Cyclic cohomology and noncommutative differential geometry.	879
DAVID, G., Opérateurs de Calderón-Zygmund.	890
DAVIE, A. M., Singular minimizers in the calculus of variations.	900
EFFROS, E. G., Advances in quantized functional analysis.	906
GARNETT, J. B., Corona problems, interpolation problems, and inhomogeneous Cauchy-Riemann equations.	917
ГЛУСКИН, Е. Д. (Paper not read at the Congress.), Вероятность в геометрии банаховых пространств.	924

HAAGERUP, U., The classification of hyperfinite von Neumann algebras. (No manuscript received.)	
JONES, V. F. R., Subfactors of type II_1 factors and related topics.....	939
KENIG, C. E., Carleman estimates, uniform Sobolev inequalities for second-order differential operators, and unique continuation theorems..	948
MILMAN, V. D., The concentration phenomenon and linear structure of finite-dimensional normed spaces.....	961
ОЛЕВСКИЙ, А. М., Гомеоморфизмы окружности, модификации функ- ций и ряды Фурье. (Homeomorphisms of the circle, modifications of functions and Fourier series.).....	976
WOLFF, T. H., Generalizations of Fatou's theorem.....	990

There were 114 Short Communications in this section.

Section 10

ARAK, T., A class of Markov fields with finite range.	994
DAVIS, M. H. A., Nonlinear filtering and stochastic flows.	1000
HOLEVO, A. S., Conditionally positive definite functions in quantum probability.	1011
KUNITA, H., Stochastic flows and stochastic partial differential equations.	1021
LIGGETT, T. M., Spatial stochastic growth models—Survival and critical behavior.	1032
PAPANICOLAOU, G. C., Wave propagation and heat conduction in ran- dom media.....	1042
STONE, C. J., A nonparametric framework for statistical modelling.	1052

There were 38 Short Communications in this section.

Section 11

ALT, H. W., Minimum problems with free boundary. (No manuscript received.)	
DIPERNA, R. J., Compactness of solutions to nonlinear PDE.....	1057
EVANS, L. C., Quasiconvexity and partial regularity in the calculus of variations.	1064
GIAQUINTA, M., The problem of the regularity of minimizers.	1072
HAMILTON, R., Parabolic equations in differential geometry. (No manu- script received.)	
IVRII, V. (Paper read by L. Hörmander.), Estimates for the number of negative eigenvalues of the Schrödinger operator with singular poten- tials.	1084
JOST, J., Two-dimensional geometric variational problems.....	1094
KRYLOV, N. V., Some new results in the theory of nonlinear elliptic and parabolic equations.....	1101
SCHEFFER, V., A self-focusing solution to the Navier-Stokes equations with a speed-reducing external force.	1110
SOLONNIKOV, V. A., Free boundary problems and problems in noncom- pact domains for the Navier-Stokes equations.	1113

TAUBES, C. H., Gauge theories and nonlinear partial differential equations.....	1123
TROMBA, A. J., Recent progress on the classical problem of plateau. ...	1133
URAL'TSEVA, N. N., Estimates of derivatives of solutions of elliptic and parabolic inequalities.....	1143
There were 35 Short Communications in this section.	

Section 12

JAKOBSON, M. V. (Paper read by J. Guckenheimer.), Families of one-dimensional maps and nearby diffeomorphisms.....	1150
KOZLOV, V. V., Phenomena of nonintegrability in Hamiltonian systems.	1161
KRYAZHIMSKII, A. V., Optimization of the ensured result for the dynamical systems.	1171
KUPKA, I. A. K., Generalized Hamiltonians and optimal control: A geometric study of the extremals.	1180
MATHER, J. N., Dynamics of area preserving maps.	1190
PESIN, YA. (Paper not read at the Congress.), Ergodic properties and dimensionlike characteristics of strange attractors that are close to hyperbolic.	1195
SERIES, C., Symbolic dynamics for geodesic flows.	1210
SULLIVAN, D., Quasiconformal homeomorphisms in dynamics, topology, and geometry.	1216
TAKENS, F., Homoclinic bifurcations.	1229
ZEHNDER, E., The Arnold conjecture for fixed points of symplectic mappings and periodic solutions of Hamiltonian systems.....	1237
ZIMMER, R. J., Actions of semisimple groups and discrete subgroups. ..	1247

There were 43 Short Communications in this section.

Section 13

BALABAN, T., Ultraviolet stability problems in quantum field theories. .	1259
ECKMANN, J.-P., The mechanism of Feigenbaum universality.	1263
GALLAVOTTI, G., Renormalization theory and group in mathematical physics.	1268
GAWĘDZKI, K., Renormalization: From magic to mathematics.	1278
MANIN, YU. I. (Paper read by I. Singer.), Quantum strings and algebraic curves.	1286
PASTUR, L. A. (Paper read by T. Spencer.), Spectral properties of metrically transitive operators and related problems.	1296
SPENCER, T., Random and quasiperiodic Schrödinger operators.	1312

There were 46 Short Communications in this section.

Section 14

BRANDT, A., Multilevel approaches to large scale problems.....	1319
BREZZI, F., New applications of mixed finite element methods.	1335
CHORIN, A. J., Vortex methods and turbulence theory.....	1348

DAHLQUIST, G. G., Some questions related to numerical methods for stiff nonlinear O.D.E.'s (No manuscript received.)	
ГОДУНОВ, С. К., Проблема гарантированной точности в численных методах линейной алгебры. (Problèmes de l'exactitude garantie dans les méthodes numériques de l'algèbre linéaire.)	1353
HEJHAL, D. A., Zeros of Epstein zeta functions and supercomputers....	1362
LANFORD, O. E., Computer-assisted proofs in analysis.....	1385
ORSZAG, S. A. AND YAKHOT, V., Renormalization group analysis of turbulence.	1395

There were 33 Short Communications in this section.

Section 15

BECK, J., Uniformity and irregularity.	1400
BJÖRNER, A., Face numbers of complexes and polytopes.	1408
ЕГОРЫЧЕВ, Г. (Paper not read at the Congress. No manuscript received.), Теория перманентов и доказательство гипотезы ван дер Вардена. (Theory of permanents and proof of van der Waerden conjecture.)	
FRANKL, P., Intersection theorems for finite sets and geometric applications.....	1419
KARMARKAR, N., A new polynomial-time algorithm for linear programming. (No manuscript received.)	
SCHRIJVER, A., Polyhedral combinatorics—Some recent developments and results.....	1431
SEYMOUR, P. D., A survey of graph minors. (No manuscript received.)	

There were 47 Short Communications in this section.

Section 16

BLUM, M., How to prove a theorem so no one else can claim it.....	1444
GRIGOR'EV, D. YU., Computational complexity in polynomial algebra.	1452
KRICHEVSKY, R. E. (Paper not read at the Congress.), Retrieval and data compression complexity.	1461
PIPPENGER, N., Reliable computation in the presence of noise.	1469
РАЗБОРОВ, А. А. (Paper read by D. Grigorev.), Нижние оценки монотонной сложности булевых функций. (Lower bounds for the monotone complexity of Boolean functions.)	1478
SHAMIR, A., The search for provably secure identification schemes.....	1488

There were 10 Short Communications in this section.

Section 17

GEMAN, S. AND GRAFFIGNE, C., Markov random field image models and their applications to computer vision.....	1496
HILDENBRAND, W., Equilibrium analysis of large economies.	1518
MERTENS, J.-F., Repeated games.....	1528

RINZEL, J., A formal classification of bursting mechanisms in excitable systems.	1578
SCHWARTZ, J. T. AND SHARIR, M., Motion planning and related geometric algorithms in robotics.	1594

There were 19 Short Communications in this section.

Section 18

БАШМАКОВА, И. Г. (Paper not read at the Congress.), Диофантовы уравнения и эволюция алгебры. (Diophantine equations and the evolution of algebra.)	1612
BOS, H. J. M., The concept of construction and the representation of curves in seventeenth-century mathematics.	1629
HAWKINS, T., Cayley's counting problem and the representation of Lie algebras.	1642
WEN-TSUN, W., Recent studies of the history of Chinese mathematics.	1657

There were 14 Short Communications in this section.

Section 19

GRABINER, J. V., The centrality of mathematics in the history of western thought.	1668
KAHANE, J.-P., Enseignement mathématique, ordinateurs et calembres.	1682
SEMADENI, Z., Verbal problems in arithmetic teaching.	1697

There were 29 Short Communications in this section.

Short Communications

GIVEN BY SECTIONS WITH THE AUTHORS ARRANGED ALPHABETICALLY

Section 01—Mathematical logic and foundations

BALLARD, DAVID J., <i>An algebraic perspective of the generalized diagonal lemma.</i>
BEAUDOIN, ROBERT E., <i>Strong analogues of Martin's axiom imply Axiom R.</i>
CHONG, CHI-TAT, <i>Low minimal degrees and 1-genericity.</i>
DAVIDON, WILLIAM C., <i>Near-standard analysis.</i>
EISELE, CAROLYN, <i>Peircean thought on mathematical logic.</i>
HARKLEROAD, LEON W., <i>Local effective computability of groups.</i>
HIRAI, ATSUSHIRO, <i>Foundation of projective algebraic geometry, viewed from orthepistemics.</i>
KIRMAYER, GREG, <i>A subtheory of Ackermann's set theory.</i>
KOHLMAYR, GERHARD F., <i>A negative solution of Hilbert's first problem and refutation of tertium non datur.</i>
KÖVESI, ENDRE H., <i>Introducing fundamental number theory (FNT).</i>
LIU, SHIH-CHAO, <i>Realizability S-truth and a generalised notion of ω-consistency.</i>

- MAGIDOR, MENACHEM, ROSENTHAL, JOHN, and RUBIN, MATTIYAHU, *Some highly undecidable lattices.*
- MOSES, MICHAEL F., *Recursive discrete linear orders.*
- PERKINS, PETER, *Finite bases and computable groupoids.*
- TURQUETTE, ATWELL R., *M-valued logic with Gödel's implication.*
- VASCO, CARLOS E., *Time-dependent macro-connectives in everyday language, logic and computation.*
- WELLS, BENJAMIN, *New views of pseudorecursiveness.*
- ZHARN, P. JOSEPH, *Semiomathetic proof theory.*

Section 02—Algebra

- ALBAR, MUHAMMAD, *On Menniche groups of deficiency zero.*
- BAMBA, SIAKA KANTE, *On a Lie-algebra solvability criterion.*
- BELL, HOWARD P. and YAQUB, ADIL, *Commutativity of rings.*
- BERGER, MARC A. and FRIEDLAND, SHMUEL, *New results on composition of real quadratic forms.*
- BEYL, F. RUDOLF, *Covering groups and small presentations of the groups $SL(2, \mathbb{Z}/2^k\mathbb{Z})$.*
- BOOKER, T. HOY, *The cyclic Jacobi algorithm with an unbounded angle of rotation.*
- BOUVIER, ALAIN J., *Some remarks on Jaffard domains.*
- BRANDAL, WILLY and BARBUT, EROL, *Rings that are FGC relative to filter of ideals.*
- BUCKLEY, JOSEPH and WIEGOLD, JAMES, *Nilpotent extensions of Abelian groups.*
- CHILDS, LINDSAY N., *Taming wild extensions with Hopf algebras.*
- CLARK, JOHN, *On the commutativity of endomorphism rings of ideals.*
- COHN, LESLIE, *Lattices and orders in quaternion algebras.*
- COMER, STEPHEN D., *The lattice of conjugacy relations on a group.*
- COMERFORD, LEO P., JR., WESTON, KENNETH W., KENNE, PETER, and NEWMAN, MICHAEL F., *Verbal embeddings into groups.*
- COOKECHPOUR, A. A., *Algebraic and topological properties of the set of maximal right ideals of a ring.*
- CRAPO, HENRY, *Geometric homology.*
- DARAFSHEH, M. R., *Characters of the group $GL_6(2)$ and its automorphism group.*
- DAYTON, BARRY H., *The Picard group of a G-algebra.*
- DEACONESCU, M. and CIOBANU, GABRIEL, *Arithmetical lattice-ordered semigroups.*
- DEMARR, RALPH and GIBSON, PETER M., *On doubly stochastic matrices.*
- DIMITRIĆ, RADOSLAV M., *On products in module categories.*
- DJOKOVIĆ, DRAGOMIR Ž., *Coxeter type presentations of some sporadic finite simple groups.*
- ELLIS, GRAHAM J., *Higher dimensional crossed modules and multi-relative algebraic K-theory.*
- ERDOGDU, VAHAP, *Regular multiplication rings.*

- FINE, BENJAMIN, HOWIE, JAMES, and ROSENBERGER, GERHARD, *A Freiheitssatz for one-relator products of cyclics.*
- FONTES DE SOUSA, MARIA DE FAIMA, *On the E-optimality of some incomplete block designs.*
- GONCHIGDORZH, R., *The r.p.p. rings and reduced hereditary rings.*
- GELLER, SUSAN C., *On the K-theory of the cusp.*
- HAILAT, MOHAMMAD Q., *Classifications of rank 2 GW-rootsystems.*
- IARROBINO, ANTHONY A., *Associated graded algebra of a Gorenstein Artin algebra.*
- JAFARIAN, ALI-AKBAR and SOUROUR, A. R., *Invertibility-preserving linear maps on the space of self-adjoint matrices.*
- JAMBOR, PAUL E., *Onesided splitting ring extensions: A generalization of Ore extensions.*
- JENDA, OVERTOUN M. G., *Modules over commutative noetherian rings.*
- JOHNSON, STEVEN D., *Computer-generated resolutions.*
- KALMAN, JOHN A., *Lattices of polynomials under substitution.*
- KEGEL, OTTO H. and SALECKER, OTTO, *Relatively lawless ordered groups.*
- KOSLOWSKI, JURGEN, *Completions of categories.*
- KROP, LEONARD, *Associated graded algebra of a relatively free group.*
- KUKU, ADEREMI O., *Some finiteness results in the higher K-theory of orders and group-rings.*
- LANG, JEFFREY and BLASS, PIOTR, *The effect of generic conditions on the factoriality of Zariski rings.*
- LEHRER-ILAMED, YEHIEL, *On explicit characteristic polynomials for central polynomial rings.*
- MACDONALD, IAN D., *The latest thing in nilpotent quotient algorithms.*
- MAJUMDAR, SUBRATA, *Cohomology and polycyclic groups.*
- MAKAR-LIMANOV, LEONID, and MALCOLMSON, PETER, *Free subalgebras of enveloping fields.*
- MALCOLMSON, PETER, *Matrix localizations of Hermite rings.*
- MCCAUGHAN, DENNIS J., *Commutator properties and subnormality.*
- МЕКЕЇ, А. (МЕКЕЇ, А.), *О подалгебрах конечной коразмерности. (On subalgebras of finite codimension.)*
- MICHAELIS, WALTER J., *Properness of Lie algebras and enveloping algebras.*
- MOGHADDAM, MOHAMMAD REZA R. and AMBERG, B., *Products of locally finite groups with min-p.*
- MYUNG, HYU C. and SAGLE, ARTHUR A., *On the construction of reductive Lie-admissible algebras.*
- NAGAHARA, TAKASHI, *Splitting rings of separable skew polynomials.*
- OLSON, DWIGHT M. and VELDSMAN, STEFAN, *An elementwise characterization of the US-prime radical.*
- OMAR, ABDEL-RAOUF A. HUSSEIN, *The superimprimitive subgroups.*
- PENG, TSU A., *The centre of a finite group.*
- PICCARD, SOPHIE, *Automorphisms and bases of finite groups of rank 2: Applications to the symmetrical and alternating groups.*
- PILZ, GÜNTER F., *Near-rings, automata, and discrete systems.*

- RATCLIFFE, JOHN G., *A uniqueness theorem for amalgamated products of groups.*
- REN, HONGSU, WAN, ZHEXIAN and WU, XIAOLONG, *Automorphisms of $PSL(2, K)$ over skew fields.*
- ROSENKNOP, JOHN Z., *Monomial filtration of modules over polynomial rings.*
- SATO, MASAHISA, *On finite dimensional algebras over which every indecomposable module has different composition factors.*
- SCHAFER, RICHARD D., *Structurable algebras and bimodules.*
- SHARAFEDDIN, AHMAD, *Sur quelques propriétés de la fonction.*
- SIBLEY, THOMAS Q., *Generalizing groups geometrically.*
- SPAKE, REUBEN, *Notes on the power semigroup of the group of integers.*
- TAMARI, DOV, *Associativity theory and the 4-color-map problem; word-chains, polyhedra, the V_1 language, lists, and linear algebra over $GF(3)$.*
- THOMSEN, MOMME JOHS, *Embeddings of near-algebras.*
- VAN GEEL, JAN, *Cancellation property for maximal orders in quaternion algebras over global function fields.*
- VUKOMANOVIĆ, DJORDJE V., *1-deductiveness and non-associative commutative residuated groupoids.*
- WERDIER, DIETER, *On old problems concerning nil structures in associative rings.*
- WIID, FRANS G., *Pole assignability of rings of low dimension.*
- YAMAOKA, KENYA, *On a problem of the power semigroup of a finite group.*
- ZOKAYI, ALIREZA, *A characterization of the groups $U_6(q^2)$.*

Section 03—Number theory

- ADLEMAN, L. M., ESTES, D., and MCCURLEY, K. S., *Solving bivariate quadratic congruences in random polynomial time.*
- AGOU, SIMON, *Polynômes sur les corps finis de Galois.*
- ALEMU, YISMAW, *On zeros of diagonal forms over p -adic fields.*
- ALEX, LEO J. and FOSTER, LORRAINE L., *On the Diophantine equation $1 + x + y = z$.*
- ALLADI, KRISHNASWAMI, *Brun's sieve and bounded multiplicative functions.*
- ANAND, BHUPINDER SINGH, *A minimal prime generating algorithm.*
- APOSTOL, TOM M., *Analytic continuation of a function of Ramanujan.*
- ARPAIA, PASQUALE J., *A report on prime prime numerals.*
- BAICA, MALVINA, *More units from the periodicity of an algorithm (Preliminary report).*
- BARNES, FRANK W., *The Diophantine equation $x^2 + 2 = y^p$.*
- BLECKSMITH, RICHARD, BRILLHART, JOHN, and GERST, IRVING, *Parity function parity theorems and theta function identity analogues.*
- CAYFORD, AFTON H., *Differential dependence of entire functions of a non-Archimedean variable (Preliminary report).*
- DASHDORZH, TS. (ДАШДОРЖ, Ц.), *Об одном обобщении теоремы Шюра. (On a generalization of Schur's theorem.)*
- DOUGHTY, HERB, *Finite double fields.*

- EVERTSE, JAN-HENDRIK and GYÖRY, KÁLMÁN, *Finiteness criteria for decomposable form equations.*
- FORSYTHE, W. L., *Chinese remainder theorem extensions to Z_M modules, elementary algebraic codes and genus 0 curves.*
- FROEMKE, JON and GROSSMAN, JERROLD W., *An algebraic and combinatorial approach to some number theoretic problems of Hilton and Pedersen.*
- GERTH, FRANK E., III, *Densities for certain p -ranks in cyclic fields of degree p^n .*
- HAN, SANG-GEUN, *p -adic L -functions and Riemann-Hurwitz genus formula.*
- HEMARD, DENIS and NGUYEN QUANG DO, THONG, *Determining the first layers of Z_p -extensions of an algebraic number field.*
- HENSLEY, DOUGLAS, *The distribution of round numbers.*
- KUIPERS, L. and SHIUE, PETER J.S., *On the L^p -discrepancy of certain sequences.*
- KUNOFF, SHARON M., *$N!$ is Benford.*
- LAMB, JOHN F., JR., *Factorialsum number chains.*
- LANGEVIN, MICHEL and LOUBOUTIN, ROLAND, *Problèmes Diophantiens.*
- LBEKKOURI, ABOUBAKR, *On the construction of a Hilbert field.*
- MAWYER, FARLEY, *Units in parametrized p -adic number fields.*
- MOLLIN, RICHARD A., *Class numbers and Diophantine equations.*
- MRCHEV, SIMEON J., *Mathematical models of the problems for optimal distribution of the neuro-ensembles and nets of links.*
- MUSES, C., *Hypernumbers: Their algebras and geometries.*
- RIESEL, HANS, *Some soluble cases of the discrete logarithm problem.*
- ROESLER, FRIEDRICH, *Riemann's hypothesis as an eigenvalue problem.*
- SCHOEN, ALAN H., *Exponential sum patterns produced by computer graphics.*
- SCHWARZ, WOLFGANG K. and SPILKER, JÜRGEN, *A variant of proof of Daboussi's theorem on the characterization of multiplicative functions with nonvoid Fourier-Bohr spectrum.*
- SMARANDACHE, FLORENTIN, *An infinity of unsolved problems concerning a function in number theory.*
- SPIRO, CLAUDIA A., *Local distribution of values of the group-counting function.*
- STEIN, SHERMAN K., *A number-theoretic problem suggested by tiling R^n .*
- TICHY, ROBERT F., *A metric estimate for the discrepancy of a special sequence.*
- YATES, SAMUEL, *Smith numbers.*

Section 04—Geometry

- ADAMS, MARK S., *Absolute volumes of the tenth and thirteenth Archimedean solids.*
- AVILES, PATRICIO U. and MCOWEN, ROBERT, *Recent developments in conformal deformations on non-compact Riemannian manifolds.*
- AWAKEEM, NOURIA A., *On incidence groups.*

- BERESINA, LYUBA J. and BEREZINA, MIRJAM T., *Affine and projective transformations of finite planes.*
- CECIL, THOMAS E., *Taut and Dupin hypersurfaces in Lie geometry.*
- CHAN, JEAN B., *Arcwise convexity of the p -kernel.*
- DEKSTER, BORIS V., *On reduced convex bodies.*
- DEL RIEGO, LILIA, *Lie subalgebras of vector fields over a differentiable manifold.*
- DODSON, CHRISTOPHER T. J., *Connections over connections.*
- DONELAN, PETER S., *Local geometry of rigid body motions.*
- DUNHAM, DOUGLAS J., *Results on hyperbolic color symmetry.*
- ENOMOTO, KAZUYUKI, *The Gauss map of flat surfaces in R^4 .*
- FARRAN, HANI R. and ROBERTSON, STEWART A., *Integrable spreads and spaces of constant curvature.*
- FILLMORE, JAY P. and PALUSZNY, MARCO, *Apollonian problem.*
- FOSTER, LEE, *Would Leibnitz lie to you?*
- GLUCK, HERMAN, ZILLER, WOLFGANG, and MORGAN, FRANK, *Area-minimizing surfaces in Grassmannians.*
- GOLDBERG, VLADISLAV V., *Nonisoclinic 2-codimensional 4-webs $W(4, 2, 2)$ of maximum 2-rank.*
- GU, CHAOHAO, *Construction of mixed extremal surfaces in Minkowski 3-space.*
- HARRIS, STEVEN G., *Complete spacelike codimension-one immersions in Lorentz manifolds.*
- JAKUBOWICZ, ANTONI, *Classification of 4-dimensional Riemann spaces and Einstein space-times with respect to the algebraic ranks a, b, c of curvature tensor.*
- KALHOFF, FRANZ B., *On local-Archimedean orderings of projective planes.*
- KHALIFAH, MOHAMMED F., *Orthogonal-relation in affine plane.*
- KHARE, S. S., *On Dold manifolds.*
- KOCH, LISA K., *Chains and Lorentz geometry.*
- KUPERBERG, WLADZIMIERZ, *On the packing and covering densities of convex plane bodies.*
- LEICHTWEISS, KURT, *On the affine surface area of convex bodies.*
- MACDONNELL, JOSEPH F., *Self-intersections of some rational ruled surfaces.*
- MACKENZIE, KIRILL, *When is a 2-form the curvature of a connection?*
- MEYERHOFF, ROBERT, *Invariants for hyperbolic 3-manifolds.*
- MICALLEF, MARIO J., *The index of minimal surfaces, isotropic immersions and the topology of positively curved manifolds.*
- NIKITIN, ALEXANDER A., *Algorithmic problems for projective planes.*
- PAPADOPOL, PETRU, *About S. Teleman's differentiable cohomology of harmonic algebras.*
- PARKER, PHILLIP E., *Sectional curvature of zonal 3-manifolds.*
- PRAKASH, NIRMALA, *On existence of ghost fields in Lorentzian manifolds.*
- SAAD, GERHARD, *Examples of embeddable metric planes.*
- SAYRAFIEZADEH, MAHMOUD, *On sets of constant affine diameter.*
- SCHMEELK, JOHN, *Infinite dimensional Fock spaces and associated creation and annihilation operators.*

- SMART, JAMES R., *Some extensions of Morley's theorem.*
- SZABO, ZOLTAN, *Hilbert's fourth problem.*
- SZENTHE, JÁNOS, *Isometric actions with orthogonally transversal submanifolds.*
- TOOMANIAN, M., *Riemannian S-symmetric spaces which are not locally symmetric.*
- TRONG THI, DAO, *Multivarifolds and minimization problems for functionals of the multidimensional volume type.*
- VERNON, MICHEAL H., *Contact hypersurfaces of a complex hyperbolic space.*
- WRONA, WŁODZIMIERZ, *Some analogues of Schur's theorem.*
- YANG, JAE-HYUN, *Holomorphic vector bundles over complex tori.*

Section 05—Topology

- ANDERSON, D. R. and MUNKHOLM, H. J., *Bounded simple homotopy over \mathbf{R} does not coincide with proper simple homotopy.*
- APANASOV, B. N., *Hyperbolic convex hulls in several dimensions, a theorem of Sullivan, and the maximal ball covers.*
- ARKOWITZ, MARTIN, *Formal differential graded algebras and homomorphisms.*
- ARYA, S. P. and BHAMINI, M. P., *Minimal pairwise P-spaces.*
- BALL, RODERICK D., *$(n + 1)$ -axial actions and the codimension 3 knot groups.*
- BARZEGAR, HAMZEH, *Bohr syntopological compactification.*
- BEER, GERALD A. and VILLAR, LUZVIMINDA V., *On weakly Baire spaces.*
- BENTLEY, H. L., HERRLICH, H., and LOWEN-COLEBUNDERS, E., *The category of Cauchy spaces is Cartesian closed.*
- BERRICK, A. J., *Acyclic groups and Eilenberg-MacLane spaces.*
- BLEILER, STEVEN A., *The knotted surface problem.*
- BOZHÜYÜK, MEHMET E., *Introduction to Trope theory.*
- BRODERSEN, HANS and WILSON, LESLIE C., *Topologically finitely-determined germs.*
- BROWN, RICHARD A., *Generalized group presentations and formal deformations of CW complexes.*
- BROWN, RONALD and LODAY, J.-L., *Computing homotopy 3-types.*
- CHURCHARD, PETER and SPRING, DAVID, *Proper knot theory in open 3-manifolds.*
- COCHRAN, TIM D., *A cobordism classification of classical links, link groups and 3-manifolds.*
- COLLINS, PETER J., *An open question in the theory of generalised metric spaces.*
- COMFORT, W. W. and ROBERTSON, LEWIS C., *Extremal pseudocompact groups.*
- ERVYNCK, GONTRAN J., *Aspects of the frontclosure topology: A coreflection.*
- FUKUI, TOSHIZUMI, *Modified analytic trivialization (MAT) of maps.*

- GILBERT, NICK D., *On the fundamental cat^n -group of an n -cube of spaces.*
- GRACE, E. E. and VOUGHT, ELDON J., *Weakly monotone images of smooth dendroids are smooth.*
- GRACE, EDWARD E. and VOUGHT, ELDON J., *Refinable maps and θ_n -continua.*
- GRIFFITHS, H. B. and NADA, S. I., *Concrete representations of topological ordered spaces.*
- HENDERSON, JAMES P., *Countably infinite dimensional manifolds.*
- HO, CHUNG-WU, *On the periodic points of a function on a manifold.*
- JONISH, DAVID and MILLETT, KENNETH C., *Invariants for embeddings of graphs in 3 space.*
- LUUKKAINEN, JOUNI, *Lipschitz and quasiconformal deformation.*
- KACHROO, PANDIT DILARAM, *Embedding of differentiable spaces.*
- KANDIL, ALY and EL-SHAFAEE, M. E., *Regularity axioms in fuzzy topological spaces and FR_i -proximities.*
- KELLY, JOHN CLIVE, *Regularity in quasi-uniform spaces.*
- LIU, YING-MING, *Completely distributive law and fuzzy Stone-Čech compactifications.*
- MANN, BENJAMIN M., *Dyer-Lashof operations on the moduli space of instantons.*
- MARITZ, PIETER, *On almost Baire class one multifunctions and a problem of Kuratowski.*
- MARXEN, DONALD J., *Free uniform groups and embeddability of topological semigroups.*
- MASUDA, TAMEZO, *The families of pseudo-open mappings.*
- MIHALIK, MICHAEL L., *Solvable groups that are simply connected at ∞ .*
- MISRA, ARVIND K., *Piecewise order arcs in hyperspaces.*
- MOREIRA, MANUEL R. F., *The distinguished generators of $\text{Ext}_{BPBP}^1(BP, BP/I_i)$.*
- MORISUGI, KAORU, *Metastable homotopy groups of $Sp(n)$.*
- OYABU, TAKASHI, *$\text{Aut}_{\text{CAT}}(M)$ and isomorphic class. $\text{CAT} = \text{TOP}, \text{PL}, \text{Diff}$.*
- PATNAIK, SURENDRA-NATH, *Maps and homomorphisms of orbit spaces under a finite group action.*
- PAUL, DEIRDRE W., *Bounded cohomology of bounded groups.*
- PEDERSEN, ERIK KJAER, *Topological group actions of $G \times H$ on spheres.*
- PUGH, CHARLES C., *Smoothing a submanifold.*
- RASSIAS, GEORGE M., *New results and unsolved problems on the minimal number of open disks that can cover a manifold of any dimension.*
- REILLY, IVAN L., *On spaces in which every denumerable subspace is discrete.*
- REPOVŠ, DUŠAN, *Detection of singularities in dimension 3.*
- RUTTER, JOHN W., *The group of self-homotopy equivalence classes using homology decomposition.*
- SHEKOURY, RAYMOND N., *Some results on Bordism fibrations.*
- SHICK, PAUL L., *Odd primary periodic phenomena in the classical Adams spectral sequence.*

SMITHSON, R. E. and LAGRANGE, R. H., *Inverse limits and scattered spaces.*

SMOLINSKY, LAWRENCE J., *Double disk knots and doubly sliced knots.*

STARK, CHRISTOPHER W., *Limit sets, root closure, and splitting theorems in negatively curved manifolds.*

Section 06—Algebraic geometry

BUJALANCE, EMILIO and ETAYO, JOSE JAVIER, *A characterization of q -hyperelliptic compact planar Klein surfaces.*

DURFEE, ALAN H., *Recent results in singularity theory.*

ESNAULT, HÉLÈNE I., *Logarithmic De Rham complexes and vanishing theorems.*

HOYT, WILLIAM L., *Elliptic fiberings of Kummer surfaces.*

KAJI, HAJIME, *Generators of Brauer groups and rational solutions of quadratic forms.*

LITTLE, JOHN B., *On Lie's approach to translation manifolds.*

LUBKIN, SAUL, *Some recent developments of p -adic cohomology.*

MAEHARA, KAZUHISA, *A finiteness theorem of varieties of general type.*

MCBRIEN, VINCENT O. and LEARY, KEVIN J., *The hypersurface representation of a class of varieties.*

O'SHEA, DONAL B., *Pairs of limits of tangent spaces and equimultiplicity.*

ROSSELLO, FRANCISCO and XAMBO-DESCHAMPS, S., *Free Chow groups.*

TOVAR, LUIS M., *Algebraic geometry.*

TRAUTMANN, GUENTHER, *A Poncelet type theorem for hypersurfaces and vector bundles on projective space.*

WEINTRAUB, STEVEN H. and LEE, RONNIE, *The structure of Siegel modular varieties of degree two.*

ZIMMER, HORST G., *A boundedness theorem for the torsion of elliptic curves over algebraic number fields.*

Section 07—Complex analysis

ACHUTHAN, P., *On recent research advances in Padé approximants and applications.*

AL-RASHED, ABDALLAH M. and ZAHEER, NEYAMAT, *Co-converial reflector curves with applications.*

AULASKARI, RAUNO P. and LAPPAN, PETER A., *On rotation automorphic functions.*

BAYOUMI, ABOUBAKR A., *Levi problem.*

CHADEMAN, ARSALAN, *Polynomial sections of holomorphic functions.*

CHEN, LI-CHEN, *Asymptotics of certain hypergeometric polynomials and $6-j$ symbols.*

EENIGENBURG, PAUL J., *On the growth of area for subordinate functions.*

ERKAMA, TIMO, *Rational maps and symmetry.*

FRIDMAN, BUMA L., *Analytic approximations of domains in C^n .*

FURUSAWA, HARUSHI, *Kleinian group.*

- GASPER, GEORGE, *A quantum generalization of the system of differential equations used by de Branges in his proof of the Bieberbach conjecture.*
- GILLIGAN, BRUCE C., *The Kobayashi pseudometric on homogeneous spaces.*
- GOODMAN, A. W., *On the zeros of the derivative of a rational function.*
- GROSS, FRED and OSGOOD, C. F., *Fixed points of composite meromorphic functions.*
- HASLINGER, FRIEDRICH, *Weighted spaces of entire functions.*
- HILL, C. DENSON, *What is the notion of a complex manifold with a smooth boundary?*
- HOOD, RODNEY TABER, *Euler's constant.*
- IWANIEC, TADEUSZ and NOLDER, CRAIG A., *Hardy-Littlewood inequality for quasiregular mappings in certain domains in R^n .*
- JIANG, CHUN-XUAN, *Theory of new complex variables.*
- KASANA, HARVIR S., *The construction and approximation of proximate orders.*
- KOÇAK, CEVDET, *An extremum problem in conformal mapping.*
- LEBRUN, CLAUDE R., *Thickenings of complex manifolds.*
- LIEB, INGO and RANGE, R. MICHAEL, *The kernel of the $\bar{\partial}$ -Neumann operator.*
- MARTIN, E. DALE, *Functions of three-dimensional complex variables.*
- MATSUOKA, CHO-ICHIRO, *Jensen measures and maximal functions of uniform algebras.*
- MILLER, SANFORD S. and MOCANU, PETRU C., *Differential subordinations and inequalities in the complex plane.*
- MOHSENI, MAHMOUD, *Finding the zeros of an analytic function by homotopy method.*
- NAFTALEVICH, AARON, *Power series and their real parts.*
- OSGOOD, CHARLES F., *Proof of the n small function theorem.*
- OWA, SHIGEYOSHI, *Generalization theorems for convolution and Komatu's conjectures.*
- PEROVIĆ, MIODRAG M., *Boundary correspondence of automorphisms of a ball with unbounded coefficient of quasiconformality.*
- RASSIAS, THEMISTOCLES M., *New inequalities for complex-valued polynomial functions.*
- SAITOH, SABUROU, *Fourier-Laplace transforms, the Bergman spaces and Paley-Wiener's theorem.*
- SIGURDSSON, RAGNAR, *Growth properties of analytic and plurisubharmonic functions of finite order.*
- SINGH, RAM, *Linear sums of certain analytic functions.*
- SMITH, J. C., *Spaces having closure-preserving covers by nice sets.*
- SONI, KUSUM K., *An extension of the Laurent series expansion.*
- ULUÇAY, CENGİZ, *On a lemma of Shaffer-Spencer.*
- UY, NGUYEN X., *Removable singularities of analytic and harmonic functions.*
- VUORINEN, MATTI, *Quasiconformal distortion in R^n .*
- WADLEIGH, FRANK R., *Generalized zeta-functions.*

WEN, GUO-CHUN, *Applications of complex analysis to nonlinear elliptic systems of partial differential equations.*

YAMASHITA, SHINJI, *Functions of uniformly bounded characteristic.*

YANG, CHUNG-CHUN, *Factorization theory of meromorphic functions.*

Section 08—Lie groups and representations

BERGMAN, GEORGE MARK, *The amalgamation basis of the category of compact groups.*

DICKENSTEIN, ALICIA and SESSA, CARMEN, *Behaviour of nilpotent orders.*

LAI, K. F., *Tate's conjecture on surface.*

MANOCHA, H. L., *On Lie algebra of difference-differential operators and hypergeometric functions.*

NIEMIEC, WACŁAW, *On new partial differential constitutive state equations of the mass/charge, energy and momentum for the mathematical physics part I: The new construction of the partial differential constitutive state equations of the mass/charge, energy and momentum.*

OYUKE, BENJAMIN CALLEB, *On the non-uniqueness of solutions for a system of equations of nonlinear elastic materials.*

PARVIZI, JAMSHID, *On local representations of Lorentz group.*

PATERSON, ALAN L. T., *An explicit formula for the Plancherel measure on nilpotent Lie groups.*

PEDERSEN, NIELS VIGAND, *Composition series of $C(G)$ and $C_c^\infty(G)$, where G is a solvable Lie group.*

PROCTOR, ROBERT A., *Odd symplectic groups and root systems of type BC .*

SCHLICHTING, GÜNTER H., *On the periodicity of group operations.*

SESAY, MOHAMED W. I., *Harmonic analysis on semisimple Lie groups.*

SIDDIQI, RAFAT N., *On density of Fourier coefficients of a function of Wiener's class.*

SNOW, DENNIS M., *Vanishing theorems on Hermitian symmetric spaces.*

SOTO-ANDRADE, JORGE, *Geometric Gel'fand models for some classical groups.*

TOWBER, JACOB, *Standard bases for representations of the classical groups.*

Section 09—Real and functional analysis

ABIAN, ALEXANDER, *Solutions of special norm of systems of infinite linear equations.*

ABIODUN, R. F. A., *Some formulae for hypergeometric series.*

APPELL, JÜRGEN, *Analyticity of nonlinear operators.*

ARAZY, JONATHAN, FISHER, STEPHEN D., and PEETRE, JAAK, *Hankel operators on weighted Bergman spaces.*

ARCHBOLD, ROBERT J., *On states of operator algebras.*

ASH, J. MARSHALL, *Concentrating trigonometric polynomial idempotents on small sets.*

ASKEY, RICHARD, *Some multiple integrals and sums.*

- AZAR PANAH, F., *On Riemann-Stieltjes integral.*
- BACHELIS, GREGORY F. and SAEKI, SADAHIRO, *Banach algebras with uncomplemented radical.*
- BALLEVÉ LANTERO, M. E. and JIMÉNEZ GUERRA, P., *On the completeness of L^α for locally convex spaces.*
- BATOR, ELIZABETH M., LEWIS, PAUL W., and RACE, DAVID, *An application of the theory of Pettis integration to operator theory.*
- BAYASGALAN, TSÈMBÈLTSGOT (БАЯСГАЛАН,ЦЭМБЭЛЦОГТ), *Числовая область линейных операторов в пространствах с индефинитной метрикой. (The numerical range of linear operators in spaces with an indefinite metric.)*
- BEER, GERALD, *Hausdorff distance and a compactness criterion for continuous functions.*
- BEG, ISMAT AND and AZAM, AKBAR, *Fixed point theorems for Kannan mappings.*
- BEHREND, EHRHARD I., *On the geometry of operator spaces.*
- BELLENGER, JOSEPH C. and SIMONS, STEPHEN, *Results on the zeros of strict multifunctions.*
- BERMAN, K. KAFTAL, V. and WEISS, GARY, *Bounded matrices with non-negative entries are paveable and related norm inequalities.*
- BIERSTEDT, KLAUS D. and BONET, JOSE, *Stefan Heinrich's density condition for Fréchet spaces.*
- BRUCKNER, ANDY M. and HU, THAKYIN, *On scrambled sets for chaotic functions.*
- BUSKES, GERARD and KRANZ, PRZEMO, *Sandwich and extension theorems in vector lattices.*
- CARTON-LEBRUN, C., *Properties of the Hilbert transform of distributions.*
- CENGİZ, BAHATTIN, *On the dual of L^1 .*
- CHENG, MIN-DE, *Fractional integral operators.*
- CHIEN, MAO-TING, *Closed derivations on the unit square.*
- CHO, TAE-GEUN, *Epimorphisms of C-algebras.*
- CIMA, JOSEPH A. and TIMONEY, RICHARD M., *The Dunford Pettis property for certain planar uniform algebras.*
- DAUGHTRY, JOHN, *Operator factorization along several nests.*
- DEEB, W., KHALIL, R. and MARZUQ, M., *Isometric multiplication of Hardy-Orlicz spaces.*
- DEGHAN, Y. N., *Directed unions of locally multiplicatively convex algebras.*
- DJEDOUR, M., *On the representation of vector valued functions in (F) and (LF) spaces by Laplace and Laplace-Stieltjes integral.*
- DUNKL, CHARLES F., *Reflection groups and orthogonal polynomials on the sphere.*
- DUYOS, SARAH M., *Universal elements in topological spaces.*
- ELLIOTT, DAVID, *Bounds for the p-norm of the finite Hilbert transform.*
- ELTON, JOHN and HILL, THEODORE P., *A generalization of Lyapounov's convexity theorem to measures with atoms.*
- ERIKSSON, SIRKKA-LIISA and LEUTWILER, HEINZ, *One application of potential theoretic methods in operator theory.*

- FAOUR, NAZIH S., *On the spectrum of weighted operator shifts.*
- FAXÉN, BIRGER J., *Approximation by equidistant translates.*
- FEICHTINGER, HANS G., *The appropriate frame for harmonic analysis over lca groups.*
- FITZPATRICK, SIMON P. and GILES, JOHN R., *The differentiability of real functions on normed linear spaces using generalized subgradients.*
- FLEISCHER, ISIDORE, *Convergence in measure for semigroup valued integrals.*
- FLEMING, RICHARD J. and JAMISON, JAMES E., *Compact operators and M-ideals (Preliminary report).*
- FLEMING, RICHARD J. and JAMISON, JAMES E., *Hermitians and isometries of a special class of Banach spaces.*
- FREIDKIN, SOLOMON, *Operators without index.*
- FRIEDMAN, YAAKOV and RUSSO, BERNARD, *The contractive projection problem and algebraic structure based on geometry.*
- FRIEDMAN, YAAKOV and RUSSO, BERNARD, *The Gelfand Naimark theorem for JB-triples.*
- FUGLEDE, BENT, *Product representation of stochastic and bistochastic matrices.*
- GERONIMO, JEFFREY S., *Ergodic theory and orthogonal polynomials.*
- GOLDSTEIN, MYRON, HAUSSMANN, WERNER, and ROGGE, LOTHAR, *Inverse mean value property of harmonic functions.*
- GOODSON, GEOFFREY R., *Spectral properties of Morse sequences over N symbols.*
- GRAF, SIEGFRIED, *Statistically self-similar fractals.*
- GROBLER, JACOBUS J., *The spectral radius of σ -order continuous band irreducible operators.*
- HAKEDA, JOSUKE, *Additivity of \ast -semigroup isomorphisms among \ast -algebras.*
- HALPERN, HERBERT, KAFTAL, VICTOR, and WEISS, GARY, *Relative Dixmier property for von Neumann algebras.*
- HALPERN, HERBERT, KAFTAL, VICTOR, and WEISS, GARY, *Matrix pavings and Laurent operators.*
- HOCHWALD, SCOTT H., *Nonlinear characterizations of the adjoint operation.*
- IFANTIS, EVANGELOS K. and SIAFARIKAS, PANAGIOTIS D., *A differential equation for the positive zeros of the function $\alpha J_\nu(z) + \gamma z J'_\nu(z)$.*
- INFANTOZZI, CARLOS, *On the series of positive terms.*
- ISAKOV, NIKOLAI M. (ИСАКОВ, НИКОЛАЙ М.) and МОУКАМБА, ФИДЕЛ' (МОУКАМБА, ФИДЕЛЬ), *Об анизотропно аппроксимативно дифференцируемых отображениях. (On anisotropically approximately differentiable mappings.)*
- IZUCHI, KEIJI and IZUCHI, YUKO, *Inner functions and division in $H^\infty + C$.*
- JIMÉNEZ GUERRA, P. and BALBAS, A., *A Radon-Nikodym theorem for a bilinear integral in locally convex spaces.*
- JIMENÉZ POZO, M. A., *Some approximation problems of geological models.*

- JULG, PIERRE and VALETTE, ALAIN J., *The Selberg principle for split rank 1 p -adic groups.*
- JURKAT, WOLFGANG and WATERMAN, DANIEL, *Conjugate functions and the Bohr-Pál theorem.*
- KAIRIES, HANS-HEINRICH, *Functional equations for nowhere differentiable functions.*
- KANNAPPAN, P., *On a generalization of the sum form-vi.*
- KHAN, ABDUL RAHIM and ROWLANDS, KEITH, *On locally solid topological lattice groups.*
- KITTANEH, FUAD A., *Inequalities for the Schatten p -norm.*
- KNOEBEL, ARTHUR, *A little bit of module theory in universal algebra.*
- KRABBE, GREGERS, *Linear operators and operational calculus, Part III.*
- KUFNER, ALOIS and OPIC, BOHUMÍR, *Compact imbeddings in weighted Sobolev spaces.*
- LAURSEN, KJELD B., *Generalized spectral capacities.*
- LÓPEZ LAGOMASINO, GUILLERMO, *Asymptotics of polynomials orthogonal with respect to varying measures.*
- LORING, TERRY A., *AF embeddings of the rational rotation algebras which are injective on K_0 .*
- LUKE, STANLEY D., *On the product of two summability methods.*
- LYONS, RUSSELL D., *Mixing and asymptotic distribution modulo 1.*
- MADAAN, V. D., *Subgroup homotopisms and their extensions.*
- MADGUEROVA, ANDREANA STEFANOVA, *On some algebras of complex-valued functions on differentiable manifolds.*
- MAHDAVI-HEZAVEHI, M., *Ordering on rings and their associated skew fields.*
- MIKOLÁS, MIKLÓS, *On a new class of nonlinear metric spaces.*
- NEL, LOUIS D., *Optimal locally convex spaces and differential calculus.*
- NEUMANN, MICHAEL M., *Decomposable operators and automatic continuity.*
- NIKNAM, ASSADOLLAH, *Automatic continuity of linear mappings in normed algebras.*
- PASCALI, D. DAN, *On variational nonlinear eigenvalue inclusions.*
- PATHAK, RAM S., *Some fractional integrals of generalized functions.*
- PETRYSHYN, W. V., *A unified theorem on the solvability of $x \in T(x)$ in cones with T k -set-contractive.*
- RADJABALIPOUR, M., *An extension of Putnam-Fuglede theorem for hyponormal operators.*
- RAO, GANDIKOTA LAKSHMI NARASINGA, *Inversion and uniqueness theorems for two distributional generalized Laplace transforms.*
- RAO, M. M., *Weakly harmonizable functions.*
- REDHEFFER, RAYMOND M., *On the maximum of the minimum of a sum.*
- RESSEL, PAUL, *Integral representations on convex semigroups.*
- RHOADES, BILLY E., *The fine spectra for weighted mean operators in $B(\ell^p)$.*
- ROSS, BERTRAM and KALLA, SHYAM, *Functional relations via fractional operators.*

- SAHAB, SALEM and KHALEELULLA, S. M., *Some results in 2-semi inner product spaces.*
- SAUER, NIKO, *Evolution equations in two Banach spaces.*
- SCHOISSENGEIER, JOHANNES, *Best possible bounds $\sum_{n \leq N} (\{n\alpha\} - 1/2)$.*
- SEDDIGHI, KARIM, *On certain invariant subspaces of operators with rich eigenvalues.*
- SIDDIQI, JAMIL A., *On the equivalence of certain Carleman classes of C^∞ -functions.*
- SINGH, DINESH and SINGH, UDIT N., *Invariant subspaces in VMOA.*
- SINGH, RAJ K., *Composition operators on weighted spaces.*
- SUNDARESAN, K. and SWAMINATHAN, S., *Gradients of convex functions in Banach lattices.*
- TAMURA, TAKAYUKI, *On the power semigroup of completely 0-simple semigroup.*
- TREIMAN, JAY, *Sharpening Clarke's generalized gradients: Optimality conditions.*
- UPMEIER, HARALD, *Index theory for multivariable Toeplitz operators.*
- VAN ELDIK, PETER and VENTER, LUCAS M., *The generalized sign-function in complex vector lattices.*
- VERMA, SADANAND, *On Fourier transforms and Fourier integrals via derivatives and their weighted jumps (Preliminary report).*
- WEIS, LUTZ, *Banach lattices with the Kadec-Pelczynski-property.*
- WOLF, EDWIN M., *Functions in $R^2(E)$ and fine interior points.*
- WU, PEI-YUAN, *Approximation by invertible and noninvertible operators.*
- XIA, DAOXING, *Trace formula for almost Lie group of operators and cyclic one-cocycles.*
- YADAV, B. S., *Approximation in weighted Banach algebras of power series.*
- YLINEN, KARI E., *An ergodic theorem for Fourier transforms of non-commutative analogues of vector measures.*
- YOSHIMOTO, TAKESHI, *A random ergodic theorem.*
- YUAN, CHUAN KUAN, *Inner invariant means and inner invariant extension of Dirac measure.*
- ZELAZKO, WIESŁAW T., *A Shilov type theorem for general topological algebras.*

Section 10—Probability and mathematical statistics

- BADAMCHIZADEH, A., *Parametric estimation in lumped Markov chains.*
- BANDELOW, CHRISTOPH, *On the arc sine law for sojourn times.*
- BENDAT, JULIUS S., *Nonlinear system analysis of a square law system with sign.*
- BRADEN, LAWRENCE S., *Probability.*
- BROOKS, JAMES K. and DINCULEANU, NICOLAE, *Optional and predictable projections of abstract stochastic processes.*
- BRUNK, HUGH D., *The compact ramification of a set, and induced measures.*

- CHANDHRY, M. L. and KASHYAP, B. R. K., *On the form of queue length and waiting time distributions in bulk queueing systems.*
- CHOU, CHING-SUNG, *Sur le difféomorphisme de l'équation différentielle stochastique du type Schwartz-Meyer.*
- DALANG, ROBERT C., *On infinite perfect graphs and randomized stopping points on the plane.*
- DANKEL, THAD G., *On the distribution of the integrated square Ornstein-Uhlenbeck process.*
- DARGAHI-NQUBARY, G. R., *On estimation of maximum intensities of the seismic and climatic loads with application to failure calculation.*
- FUNKENBUSCH, WALTER W., *Isotropic nested cycle conjecture.*
- GELMAN, ALEX, FRIEDMAN, Y., and PHADIA, E., *A goodness of fit test based on the best invariant estimator under Kolmogorov-Smirnov loss.*
- GODBOLE, ANANT P., *The minimal growth rate of partial maxima.*
- HAREL, MICHEL, *Linear serial rank statistics to test ARMA process against A.R.M.A. process.*
- HASHEMI-PARAST, S. M., *A distribution-free selection procedure for populations following a Lehman model using the indifference-zone formulation.*
- HOLTE, JOHN M., *Maximum litter-size distribution for a Galton-Watson process.*
- HWANG, TEA-YUAN, *On the distribution of two-sample student's t test for exponential distribution.*
- KAMINSKY, KENNETH S. and RHODIN, LENNART, *Maximum likelihood prediction.*
- KLASS, MICHAEL J. and HAHN, MARJORIE G., *Uniform approximation of log tail probabilities of Poissonized sums.*
- LI, ZEZHI, *An answer to Taylor's conjecture.*
- MAIBODI, H. YASSAEE, *A comparative study of estimators and statistics in multidimensional contingency tables.*
- MAMMITZSCH, VOLKER, *How to compute optimal kernel type estimators by orthogonal series expansion.*
- MANGANO, GIAN CARLO and NAFTALEVICH, AARON, *On a boundary crossing problem for certain stochastic processes.*
- MERKLE, MILAN J., *Weak convergence of measures on dual spaces.*
- MEYER, W. T. and GROMOLL, D., *Examples of complete manifolds with positive Ricci curvature.*
- NORITA, SADATAKA, *On the moving average stochastic processes in three dimensions.*
- PHILIPPOU, ANDREAS N., *Distributions of order k and longest runs with reliability applications.*
- PRATS, FRANCESC, *$C_{M,N}$ -conjugate transform.*
- RAJPUT, BALRAM S. and ROSINSKI, JAN, *Stochastic integrals relative to i.d. random measures with applications to the integral representations of i.d. processes.*
- RAMACHANDRAN, DORAISWAMY, *Certain types of disintegrations.*

- ROOS, C., VAN ZANTEN, A. J., and DAMERELL, R. M., *On some Diophantine equations related to the classification of generalised Moore geometries.*
- SIDAK, ZBYNEK, *A distribution-free discrimination procedure for circular data.*
- TATA, MAHABANOO N., *Estimating the maximum of the support of a beta distribution.*
- VO KHAC, KHOAN, *Asymptotically efficient sequence of estimators of a differentiable function of parameters.*
- WU, GUO-FA and HSU, ZHE, *A method for testing the goodness of prediction of a linear model.*
- WULFSOHN, AUBREY, *Nonlinear equations for random measures with deterministic solutions.*
- YUKICH, JOSEPH E., *Limit theorems for empirical processes indexed by functions.*

Section 11—Partial differential equations

- ADOMIAN, GEORGE, *Analytical solution of strongly nonlinear systems of partial differential equations in three space dimensions and time.*
- AL'BER, SOLOMON and AL'BER, MARK S., *Hamiltonian formalism for finite-zone solutions of non-linear equations.*
- ARKIN, JOSEPH and BERGUM, GERALD E., *Generalized recurrence formulas, II.*
- AUCHMUTY, GILES, *Variational principles for initial value problems.*
- BIRNIR, BJORN, *The structural stability of integrable nonlinear partial differential equations.*
- VAN CASTEREN, JOHANNES A., *Markov processes and fundamental solutions.*
- CHENG, KUO-SHUNG and LIN, JENN-TSANN, *On the elliptic equations $\Delta u = K(x)u^\sigma$ and $\Delta u = K(x)e^{2u}$.*
- COUTU, GERARD, *Theorems on simple separability of Laplace and Helmholtz equations in Riemannian n -space.*
- DEUMENS, ERIK and WARCHALL, HANK, *Explicit stable solitary waves in multiple dimensions.*
- DRÁBEK, PAVEL and KUČERA, MILAN, *Reaction-diffusion systems with unilateral boundary conditions.*
- DUFF, GEORGE F. D., *Estimates and inequalities for the Navier Stokes equations in three space dimensions.*
- ELCRAT, ALAN R. and LANCASTER, KIRK E., *Some special nonparametric minimal surfaces.*
- GARCIA, MARIANO, *K -fold isotopic amicable numbers.*
- GRÜTER, MICHAEL, *Minimal surfaces with free boundaries, and applications.*
- HORNUNG, ULRICH and JÄGER, WILLY, *A new model for chemical reaction in porous media.*
- KUNG-CHING, CHANG, *A minimax theorem without compactness and its application to the calculus of variations.*

- KUO, TSANG-HAI, *On some existence theorems for quasilinear elliptic equations.*
- LENHART, SUZANNE M., *Viscosity solutions associated with switching control problems for piecewise-deterministic processes.*
- LIEBERMAN, GARY M., *Classical solvability of elliptic equations with discontinuous boundary conditions.*
- MAMOURIAN, AHMAD, *On a boundary values problem for nonuniformly Lavrent'ev type equations.*
- MAUMARY, SERGE, *Foundations of the separation of variables.*
- NICOLOSI, F., *Boundary value problems for degenerate parabolic operators in open unbounded sets.*
- OHTSUKA, KOHJI, *A basic theory of generalized J-integral.*
- PALAIS, BOB, *Blow up for a nonlinear evolution equation and self-reproducing strained shear layers.*
- PELCZAR, ANDRZEJ, *Wazewski topological method applied to second order partial differential equations.*
- REDLINGER, REINHARD, *Parabolische systeme mit funktionalen.*
- ROȘCULEȚ, MARCEL N., *Analyticité hyperbolique.*
- SASTRY, D. R. K., *Qualitative theory of reaction diffusion equations occurring in neuro-mathematics.*
- SAXTON, RALPH A., *Membrane dynamics.*
- SPECK, FRANK-OLME, *On the quarter-plane problem in diffraction theory.*
- VAN, TRAN DUC, *Some problems of the theory of differential equations of infinite order and applications.*
- WEN, SHIH-LIANG and CHEN, YUNKAI, *Periodic wave solutions to the Kadam-tsev-Petviashvili equation.*
- WILLIAMS, MARK, *The reflection, grazing, and gliding of singularities of solutions to semilinear boundary value problems.*
- ZAYED, AHMED and HAIMO, DEBORAH TEPPER, *Inversion of an integral transform related to a general form of heat equation.*
- ZAYED, AHMED I. and HAIMO, DEBORAH T., *Inversion of integral transforms associated with a class of perturbed heat equations.*

Section 12—Ordinary differential equations and dynamical systems

- BARTOSIEWICZ, ZBIGNIEW, *Differential equations on affine varieties.*
- BHAKTA, P. C. and DAS, N. C., *Eigenfunction expansion method to thermoelastic and magneto-thermoelastic problems.*
- BURTON, T. A. and HATVANI, L., *A non-generalization of Liapunov's direct method.*
- CHEN, DANIEL K. C., *Non-classical complex numbers III: Vector division.*
- CHICHESTER, FREDERICK D., *Application of direct matrix products to the synthesis of reduced order observers for a state variable model of a flexible robot arm.*

- CLARK, JACK E., *A numerical conjugacy invariant for dynamical systems.*
- COOKE, KENNETH L. and WIENER, JOSEPH, *Differential equations with piecewise continuous arguments* (EPCA)
- DAHIYA, R. S., *On oscillation of solutions of functional differential equation with mixed arguments.*
- ELIAŠ, JOZEF, *Application of the generalized N. Aronszajn's theorem on the functional system of differential equations of n th order, $n \geq 2$.*
- EL-SABAA, F. M. F. and SHERIF, H. H., *About the periodic orbits of galactic motion.*
- ERBLAND, JOHN P. TROYER, STEPHANIE F., and WILLIAMS, JOHN, *Chaotic maps on tori.*
- FAURE, ROBERT, *Cycles d'équations différentielles non linéaires; liens entre paramètres et pulsations des solutions périodiques.*
- GINGOLD, HARRY and HSIEH, PO-FANG, *On asymptotic integration with a turning point at infinity.*
- GOPINATH, LAHA, *Transforms induced by ordinary even-order differential equations.*
- GRABOWSKI, PIOTR L., *Stability of abstract semilinear systems in Hilbert space.*
- GREGUŠ, MICHAL, *On a boundary value problem of the third order with two parameters.*
- GUTOWITZ, HOWARD VICTOR, JON, and KNIGHT, BRUCE, *A local structure theory for cellular automata.*
- HADID, SAMIR B. and ALSHAMANI, JOSEPH G., *Asymptotic solutions.*
- HAMAD, GOWEIDA DOMA, *On some new discrete inequalities involving the shift operator E .*
- HARRISON, JENNY, *Dynamical properties of continued fractions.*
- HONARY, BAHMAN K., *S^1 -equivariant dynamical systems on compact 2-manifolds.*
- HUANG, QICHANG and WANG, KE, *Generalized phase space for retarded equations with infinite delay.*
- JONES, JOHN, JR., *Nonlinear matrix differential equations.*
- KOČAK, H., *Topology and mechanics with computer graphics.*
- KOSMALA, WITOLD A., *On positive solutions of equations with two middle terms.*
- KRAIROJANANAN, S., *Application of difference equations to the Fibonacci sequence.*
- KRATZ, WERNER, *An oscillation theorem for self-adjoint Hamiltonian systems.*
- LIN, ZHEN-SHENG, *C^r closing lemma.*
- LIU, SHIZE, *On the dynamical system with some purely imaginary eigenvalues.*
- MANOLOV, SPAS M., *Systèmes autonomes à espaces des phases symétriques et des trajectoires périodiques.*
- MAYER, DAVID, *Outline of an analytic exterior algebra for the integration of curvature.*

- MEHRI, BAHMAN, *Existence of periodic solution for certain non-linear differential equation.*
- MEYER, KENNETH R. and SELL, GEORGE, *Homoclinic phenomena in almost periodic systems.*
- MORALES, PEDRO, *On semilinear evolution equations in Fréchet spaces.*
- NISHIMOTO, KATSUYUKI, *Fractional calculus and its application to non-homogeneous Gauss' equation.*
- PRANGER, WALTER A., *A new method of solution for the inverse Sturm-Liouville problem.*
- PUTTASWAMY, T. K., *Asymptotic solution of a certain differential equation near an irregular singular point.*
- REZA, F. M., *Canonic synthesis of linear passive normal RLC n-ports.*
- SHAHSHAHANI, S. M., *Simultaneous Hopf bifurcation in quadratic systems.*
- SHANTAO, LIAO, *On stability of 3-dimensional vector fields.*
- TIAN, JINGHUANG, *Hilbert's 16th problem.*
- ULLMAN, DANIEL, *H-recurrence of integrable functions.*
- ZHENG, QUAN and GALPERIN, EFIM A., *Integral global optimization method and its application to dynamical systems.*

Section 13—Mathematical physics

- AJANAPON, PIMON, *Classical limit of the path-integral formulation of quantum mechanics in terms of the density matrix.*
- ARIMA, SATOSHI and ASAEDA, YOU, *Pseudo-vectors are anti-symmetric tensors.*
- ÁVILA, GERALDO S. S., *Alfvén-type wave propagation in magnetodynamics.*
- BREEN, STEPHEN A., *Large order estimates for ground state energy perturbation series.*
- CACUCI, D. G., PEREZ, R. B., and PROTOPODESCU, V., *Duals and propagators: A new canonical formalism for general nonlinear equations.*
- COHOON, DAVID K., *Uniqueness of the solution of the transmission problem for 7 dimensional electromagnetic interaction problems.*
- DEBNATH, LOKENATH, *Nonlinear surface waves in a dissipative fluid medium.*
- DEUMENS, ERIK and WARCHALL, HENRY, *Explicit construction of all spherically symmetric solitary waves for a nonlinear wave equation in multiple dimensions.*
- DÖKMECI, CENGİZ M. and SARIGÜL, NESRİN, *Certain integral and differential types of variational principles of mathematical physics.*
- DWIVEDI, B. N., *On some magnetohydrostatic configurations.*
- FABRIKANT, VALERY I., *New results in potential theory and their applications.*
- FLAHERTY, FRANCIS J., *Yang-Mills equations on the multiplicative 4-torus*
- GHOSH, P. R., *Stress concentration around a small inclusion on the axis of a circular cylinder under torsion.*

- GOLBABAEI, A., *Phase-winding solutions for axisymmetric convection between rotating planes uniformly heated from below.*
- GRAVEL, PIERRE and GAUTHIER, CLAUDE, *Multiconnected vacua in Kaluza-Klein theories.*
- HAYASHI, HIROSHI, *Solution of Navier-Stokes equation.*
- HURD, ALBERT E., *On global existence and validity for the BBGKY hierarchy.*
- HWANG, J. S., *On balance theory in celestial mechanics and quantum theory.*
- ILLNER, REINHARD, *The Boltzmann equation.*
- KAMMERER, JEAN-BERTRAND, *Analysis of the Moyal product in a flat space.*
- KOBAYASHI, YUJI and NAGAMACHI, SHIGEAKI, *Generalized superspaces.*
- KOILLER, JAIR, BALTHAZAR, J. M., and YOKOYAMA, T., *Relaxation chaos.*
- KOSACHEVSKAYA, ELENA A. (КОСАЧЕВСКАЯ, ЕЛЕНА А.) and KOSACHEVSKIĬ, LEONID R. (КОСАЧЕВСКИЙ, ЛЕОНИД Р.), *О приближенном решении краевой задачи фильтрации неньютоновской жидкости с движущейся границей при наличии температурной зависимости реологических параметров. (On the approximate solution of the boundary value problem of filtration of a non-newtonian fluid with a moving boundary, in the presence of temperature dependence of the rheological parameters.)*
- LAPIDUS, MICHEL L., *The Feynmann-Kac formula with a Lebesgue-Stieltjes measure and Feynman's operational calculus.*
- LASSNER, GERD, *KMS-states on quasi-algebras.*
- MAHALANABIS, R. K., *Coupled dynamical problem of thermoelasticity in a non-simple elastic medium.*
- MAHMOUD, F. F., EL SHARKAWY, A., and ABDEL RAHMAN, O. E., *Optimum design of contact surfaces by the direct automated method.*
- MAKKY, SADIA M., *Design of three dimensional electrodes for any pre-assigned field distribution at the discharged midplane.*
- MARATHE, KISHORE B., *Geometry of coupled instantons.*
- MAURO, FABRIZIO, *Stability conditions in linear viscoelasticity.*
- MEJIA, RAYMOND, STAR, ROBERT A., and KNEPPER, MARK A., *Flow in a leaky tube.*
- NAGAYAMA, HARUYA, *A theory of general relativity by Otsuki connections.*
- NASIR, MOHAMMAD YAQUB, *Equation of state of the system of magneto-acoustic waves in an ideal MHD fluid.*
- NASSAR, M., *Viscoelastic plate resting on fluid.*
- PHOOLAN, PRASAD, *Kinematics of a curved nonlinear wave front and a shock front: Extension of Huyghen's method of construction of a wave front.*
- QICUN, SUN, *Physical groups and new geometries of space-time I.*
- RAHIMI-ARDABILI, M. Y., *The equilibrium of a non-uniform rotating body of arbitrary density.*
- RAWNSLEY, JOHN H., *Twistor methods.*
- ROSENBERG, STEVEN, *Determinants of first order geometric operators.*

- SAGLE, ARTHUR A. and MYUNG, HYO, *Nonassociative algebras and Lagrangian mechanics on homogeneous spaces.*
- SALZER, HERBERT E., *Some aspects of momentum research.*
- SIEGEL, E. D., *Fractals, diffractals and irrational numbers; functals: another way to do physics.*
- SLABY, MARY ANN, *The fundamental structure of the real and imaginary quaternary Khadaffi field.*
- STANKOVIĆ, BOGOLJUB, *S-asymptotic expansion of distributions.*
- TSOU, SHEUNG TSUN, *A non-Abelian gauge theoretical generalization of the Poincaré lemma.*
- VERMA, ALAKH P., *Instabilities in displacement processes in porous media.*

Section 14—Numerical methods and computing

- BEHFOROOZ, G. H., *Superconvergence properties of quintic interpolatory splines.*
- BURKHART, RICHARD H., *The discrete Green's function for a fast 3-D free-space Poisson solver.*
- CALDWELL, JAMES, *Variational iterative schemes.*
- CANUTO, CLAUDIO, *Analysis of spectral methods for the Navier-Stokes equations.*
- CUYT, ANNIE A. M., *Theory and applications of multivariate rational interpolants.*
- DE WILDE, P., EDWARDS, S. P., and DE MEYER, K., *Error estimates for a finite element solution of the PDE's modeling semiconductor devices.*
- DERR, LEROY, OUTLAW, CURTIS, and SARAFYAN, DIRAN, *Continuous solution of $y'' = f(x, y, y')$ and $y' = f(x, y)$ initial value problems.*
- DRIESSEL, KENNETH R., *An isospectral gradient flow.*
- EISSA, RAOUF PETRO, *The approximation of the n-dimensional singular integral.*
- ERXIONG, JIANG, *The modified Rayleigh quotient iteration.*
- FIEDLER, MIROSLAV, *Recent results on Bézout matrices.*
- GALAJDA, PAVEL, *Numerical and graphical methods for solution of linear and non-linear differential equations and of linear and non-linear programming problems.*
- KAVIANPOUR, A. and BAYAT, G., *Automatic design of printed circuit boards.*
- KHALIQ, A. Q. M., *Splitting methods for multi-dimensional parabolic equations.*
- LAPIDUS, ARNOLD, *Computation of radially symmetric shocked flows.*
- MALEKEJAD, K., *The applications of generalized cross-validation in deconvolution problem.*
- MCCARTIN, BRIAN J. and LABARRE, ROBERT E., *Numerical computation on arbitrary lattices of points.*
- MÜHLBACH, GÜNTER W., *Elimination strategies and applications.*

- NASH, DAVID H., *A "three returns six" algorithm for pairwise independent normal deviates.*
- NATVIG, JON, *Efficient use of the CYBER 205 for matrix multiplication.*
- NWANNENNA, CHRIS C., *An interactive system for optimising trim loss in paper manufacture.*
- PENCE, DENNIS D., *Multiplicative factorization of Lagrange splines.*
- POPOV, BLAGOJ S., *Accelerations of the convergence of certain class of sequences.*
- SARAFYAN, DIRAN, *Continuous approximate solution of ordinary differential equations and their systems.*
- SHELLEY, MICHAEL J. and BAKER, GREGORY R., *On the motion of vortex layers.*
- STRANG, GILBERT, *An algorithm for Toeplitz matrix calculations.*
- STYS, TADEUSZ, *A semifinite difference approximation of nonlinear parabolic-elliptic equations.*
- SURI, MANIL, *An optimal error estimate for the h-p version of the finite element method.*
- WANG, TSENG-CHAN, STANFORD, RICHARD H., SUNSERI, RICHARD F., and BRECKHEIMER, PETER J., *Optimal search of a spacecraft trajectory to arrive at desired targets.*
- WATHEN, ANDREW JOHN and BAINES, MICHAEL J., *Moving finite elements for the numerical solution of evolutionary partial differential equations.*
- WOOD, DICK A., *The usability of the convergence criteria for the SOR-Newton technique.*
- WOOD, GRAHAM R., *The bisection method in higher dimensions.*
- YU, WENCI, *On conjugate direction methods without computation of derivatives in nonlinear optimization.*

Section 15—Discrete mathematics and combinatorics

- AINOUCHE, AHMED, *Sufficient conditions for Hamiltonicity in graphs.*
- ALAVI, Y., CHARTRAND, G., OELLERMANN, O., CHUNG, F.R.K., ERDŐS, PAUL, and GRAHAM, R. L., *Embedding graphs in highly irregular graphs.*
- AL-THUKAIR, FAWZI A. and ZAGUIA, N., *On finding the maximum number of jumps of linear extensions of an ordered set (Preliminary report).*
- ARKIN, JOSEPH and BURR, STEFAN A., *Recurring-sequence tiling (Preliminary report).*
- AROCHA, JORGE L., *Galois connections and embedding of graphs.*
- BERGELSON, VITALY and HINDMAN, NEIL, *Combinatorially large cells of partitions.*
- CARNIELLI, WALTER A., *Ramsey-type theorems for cubes.*
- DASTRANGE, NASSER, *Various forms of sampling expansions in discrete Walsh-Fourier analysis.*
- DING, REN, KOŁODZIEJCZYK, KRZYSZTOF, and REAY, JOHN R., *A Pick-type theorem in Archimedean planar tilings.*

- DYMACEK, WAYNE M., *Steinhaus graphs*.
- GALPERIN, EFIM A., *Division-free method for linear programming: The rolling matrix algorithm*.
- GILPIN, MICHAEL J., *Powers of the stabilizing character*.
- GIMBEL, JOHN G., *The random graph as a counterexample in cochromatic theory*.
- GOODMAN, JACOB E. and POLLACK, RICHARD, *On the asymptotic number of polytopes*.
- GUSTAFSON, ROBERT A., *Hypergeometric series in $U(s)$* .
- HARARY, F. and SCHUSTER, S., *Interpolation theorems for invariants of spanning trees*.
- HARARY, FRANK, *Games with graph invariants*.
- HARBORTH, HEIKO, *Ramsey numbers for graphs with five vertices*.
- HART, SERGIU and SHARIR, MICHA, *Nonlinearity of Davenport-Schinzel sequences and of generalized path compression schemes*.
- ISMAIL, MOURAD E. H., *The variation of zeros of certain orthogonal polynomials*.
- JOHNS, GARRY and SABA, FARROKH, *On the path-chromatic number of a graph*.
- KEMNITZ, ARNFRIED, *Integral point sets*.
- KESTENBAND, B., *New families of Block designs*.
- KHELLADI, ABDELKADER, *Flows in bidirected graphs*.
- KHOSROVSHAHI, GHOLAMREZA B. and MAHMOODIAN, EBADOLLAH S., *A linear algebraic algorithm for generating a basis for null designs*.
- KHOSROVSHAHI, GHOLAMREZA B. and MAHMOODIAN, EBADOLLAH S., *On construction of BIB designs with various support sizes for $v = 9$, $k = 3$* .
- LAMKEN, ESTHER R. and VANSTONE, SCOTT A., *The existence of partitioned balanced tournament designs*.
- LAYWINE, CHARLES F. and MAYBERRY, JOHN P., *A simple construction for the two triangle-free 3-colored K_{16} 's*.
- LEE, SIN-MIN and WUI, INDRA, *On the Skolem graceful 2-stars and 3-stars*.
- LICK, DON R., *On highly regular graphs*.
- MEJTSKY, GEORGE J., *A new combinatorial method: Parallel modeling*.
- MINC, HENRYK, *Permanents of $(0,1)$ -circulants*.
- MORITZ, ROGER H. and WILLIAMS, ROBERT C., *A coin-tossing problem and some related combinatorics*.
- PISANSKI, TOMAŽ and MOHAR, BOJAN, *Embeddings of abelian groups of odd order*.
- PROPP, JAMES, *Generalized Ferrers diagrams*.
- RANDIĆ, MILAN and BAKER, BERNADETTE M., *Isospectral multigraphs*.
- RAY, NIGEL, *Umbral calculus and a generalised chromatic polynomial*.
- SINGMASTER, DAVID, *A new approach to N -person combinatorial games*.
- SNOW, DONALD R., *The generating functions for the Catalan numbers, Fibonacci numbers, and other combinatorial sequences*.
- SRIVASTAVA, JAYA, *Factorial designs based on parallel hyperplanes in Euclidean n -space based on $GF(2)$* .

- STEMBRIDGE, JOHN R., *Macdonald's conjecture for the root system A_n .*
- TALLI, MAHMOUD R., *An integer linear programming model for classifying inventory items.*
- VINCZE, ISTVAN, *An extreme value problem on permutations.*
- WHITE, ARTHUR T., *Voltage graphs and graphical products.*
- YANG, KUNG-WEI, *q -Algebras.*
- ZEILBERGER, DORON, *A quick approach to Schur functions.*
- ZOU, H. B., *Trees containing a given tree as a unique greatest common subgraph.*

Section 16—Mathematical aspects of computer science

- COHEN, MOSES E. and HUDSON, DONNA L., *The use of a new class of orthogonal functions in pattern recognition.*
- CUDIA, DENNIS F., *The role of the Weierstrass excess function in normed linear spaces and information systems.*
- GUODING, HU, *Definition of serial and parallel computation I, II.*
- HUDSON, DONNA L. and COHEN, MOSES E., *Approximate reasoning in an expert system.*
- JARVINEN, RICHARD D., *Controlling for an optimal search.*
- KAVIANPOUR, ALIREZA, *Cordic calculation.*
- KHALIMSKY, EFIM D., *Topological methods in computer science.*
- NAKASHIMA, KATSUYA, *New representation of numbers in computers.*
- QIU, HESHENG, *Polynomial algorithm in feasible basis extension problem.*
- WEYLAND, NICHOLAS J. and PUCKETT, EDWARD A., *Efficient storage of computer files of fixed and floating point numbers.*

Section 17—Applications of mathematics to nonphysical sciences

- ACZÉL, J. and ALSINA, C., *Synthesizing conditions and functions.*
- ALSINA, CLAUDI and TRILLAS, ENRIC, *On the stability of some functional equations arising in multivalued logic.*
- CHATTERJEE, GAUTAM, *A contribution to the Cambridge theories of income distribution.*
- CLARKE, A. BRUCE, *An algorithm for modifying salary regression lines.*
- CULL, PAUL, *Stability in one-dimensional population models.*
- DERRIG, RICHARD A., *Using simultaneous Ito processes to price insurance contracts.*
- DUMITRU, VINCENTIU I., *Monotone operators and Pareto equivalence in multicriteria optimization.*
- EGGHE, LEO C., *Some results on the law of Bradford.*
- GRETSKY, NEIL E. and OSTROY, JOSEPH M., *Perfect competition in large-square economies.*
- GROSS, LOUIS J., *Mathematical modeling of photosynthetic dynamics in varying light environments.*
- GUAN, PUHUA, *Cellular automata and a public key cryptographic system.*
- LEIFMAN, LEV J., KATZ, ELIMARTY, ROGER H., and ROBINSON, STEWART M., *A mathematical model of translation between natural languages.*

- LI, BING-ZHOU, *The mathematical principle of exponential smoothing.*
- OVCHINNIKOV, SERGEY V., *Representation theorems for probabilistic relations.*
- DI PASQUALE DE BORISOV, CRISTINA and SPINADEL, VERA W., *A mathematical model for the interactions among populations of living organisms.*
- SUNG, CHEN-HAN and AHANGAR, REZA, *Stochastic control modeling of some diabetes problem.*
- SWART, JOHAN, *A mathematical model to optimize culling strategies in a predation-competition farming environment.*
- VERMA, KRISHNANAND, *Electrical potential in a nerve under a certain condition.*
- ZHOU, ZHIMING, *Sign global stability for Volterra model.*

Section 18—History of mathematics

- ALTHOEN, STEVEN C. and MCLAUGHLIN, RENATE, *Will the real Jordan (in Gauss-Jordan reduction) please stand up?*
- CHOWDHURY, MUNIBUR R., *The impact of Hausdorff's "Grundzüge der Mengenlehre" on the development of point-set topology.*
- CORCORAN, JOHN, *Undefinability tests and the Erlanger Programm: Historical footnotes.*
- D'AMBROSIO, UBIRATAN, *Pre-history of calculus and its reflections in early colonial mathematics in Brazil.*
- DRUCKER, THOMAS L., *Discovery and style in Newton's early work on series.*
- FANG, JOONG, *The "Needham" question.*
- HARDING, INES, *The first mathematical publications of Chileans.*
- HARKLEROAD, LEON W. and HARKLEROAD, EDIE D., *Rózsa Péter: Recursion theory's founding mother.*
- KENSCHAF, PATRICIA CLARK, *Charlotte Angas Scott (1858 – 1931): Pioneer mathematician.*
- KELLOGG, MARY, *Mathematical methods in Isaac Newton's Philosophiæ Naturalis Principia Mathematica.*
- KREYSZIG, ERWIN, *Tschirnhaus's influence on the early development of calculus.*
- SCANLAN, MICHAEL J., *Beltrami's model and the independence of the parallel postulate.*
- SYNOWIEC, JOHN A., *Pseudo-differential operators and operational calculus.*
- TATTERSALL, JAMES J., *Waring's other problem.*

Section 19—Teaching of mathematics

- ADDA, JOSETTE, *Apprentissage de la resolution de problemes de didactique des mathematiques.*

- ALE, S. O., *Poor performance in school mathematics—Causes and remedies.*
- BEBBE-NJOH, E., *Considerations on the dialectic construction of mathematics: Its didactic implications.*
- BURGUES, CARMEN, *New materials for teaching fractions at the elementary level.*
- COONEY, MIRIAM P., *A seminar on women and mathematics.*
- DEMANA, FRANKLIN D., *Establishing fundamental concepts of algebra through numerical problem solving.*
- DURAPAU, VICTOR J. and FONTOVA, ESTHER M., *An undergraduate colloquium in mathematics and computer science.*
- FELDMAN, LEONARD, *Introductory algebra: A participatory approach for college students.*
- FJELSTAD, PAUL, *Encouraging student creativity: One result – trigonometries galore.*
- FUSARO, B. A., *The mathematical competition in modeling (MCM).*
- GALDON, JOSÉ M., *The systematic use of audiovisual teaching methods in the introduction to calculus.*
- HARTNETT, WILLIAM E., *History of modern mathematics (1850-) courses.*
- KAJIKAWA, YUJI (Presented by Hitosumatsu Shin), *Teaching by the level of achievement.*
- KATSIFLI, D. and FYFE, D. J., *Computer aided learning for numerical analysis problem solving (CALNAPS)*
- KISHORE, MASAO, *Division in decimal and binary number systems.*
- LINFOOT, JOYCE J., *Some considerations on the difficulty of questions in arithmetic.*
- LOASE, JOHN, *An alternative to computer assisted mathematical instruction – Individualized learning at Westchester Community Collge.*
- MICHAEL, MARK, *Using a microcomputer to simulate certain 2-manifolds.*
- MILES, E. P., *Printing applications for nested self-dual model subsets of the color continuum.*
- PUSTILNIK, SEYMOUR W., *Action research on reasoning and practice in mathematics achievement.*
- RAKOVER, BORIS D., *The use of algorithmic procedures in mathematical instruction.*
- REIS, M. RAQUEL, *On the integrated teaching of the different geometries.*
- RILEY, PETE EDWARD, *Teaching styles and motivation in upper division and graduate mathematical education.*
- SANATANI, SUNOY, *Can mathematics be taught?*
- THOMPSON, THOMAS M. and WIGGINS, KENNETH L., *Limits and continuity.*
- TOWERS, DAVID ANTHONY, *The role of mathematics education courses in undergraduate mathematics teaching.*
- WAITS, BERT K., *Using microcomputers to enhance the learning of pre-calculus mathematics.*
- WILSON, JOHN H., *Mathematical instruction—Focus on the development of reasoning skills.*
- YANOSKO, BARBARA J., *How long does it take to forget mathematics?*

CONGRESS PARTICIPANTS

Abacus, Alexander B., U.S.A.
 Abeasis, Silvana, Italy
 Abe, Shingo, Japan
 Abian, Alexander, U.S.A.
 Abraham, Ralph H., U.S.A.
 Abresch, Uwe, U.S.A.
 Acanfora, Fausto, Italy
 Achuthan, P., India
 Ackermans, Stan T. M.,
 Netherlands
 Acquistapace, Paolo, Italy
 Aczél, János, Canada
 Adachi, Norio, Japan
 Adams, Colin C., U.S.A.
 Adams, J. Frank, England
 Adams, Malcolm R., U.S.A.
 Adams, Mark S., U.S.A.
 Adams, Robert A., Canada
 Adams, William W., U.S.A.
 Adamson, Iain T., Scotland
 Addington, Susan L., U.S.A.
 Addison, Alonzo C., U.S.A.
 Addison, John W., U.S.A.
 Addison, West, U.S.A.
 Adeyeye, John O.-O., Nigeria
 Adleman, Len M., U.S.A.
 Adler, Roy L., U.S.A.
 Adomian, George, U.S.A.
 Aggarwal, Shiv K., U.S.A.
 Agou, Simon, France
 Ahangar, Reza, U.S.A.
 Ahdout, Shahla M., U.S.A.
 Ahlfors, Lars V., U.S.A.
 Ahuja, Mangho, U.S.A.
 Ainouche, Ahmed, Algeria
 Aissen, Michael, U.S.A.
 Aizicovici, Sergiu, Romania
 Ajanapon, Pimon, U.S.A.
 Ajetunmobi, Michael O.,
 Nigeria
 Akbulut, Selman Y., U.S.A.
 Akemann, Charles A., U.S.A.
 Akin, Kaan Y., U.S.A.
 Akis, Vladimir N., U.S.A.
 Al-Aqeel, Adnan, Kuwait
 Alavi, Yousef, U.S.A.
 Al-Bassam, M. A., England
 Albar, Muhammad A., Saudi
 Arabia
 Albar, Sahl Fadul, Saudi Arabia
 Al'ber, Solomon, U.S.S.R.
 Albers, Donald J., U.S.A.
 Albert, Douglas J., U.S.A.
 Albertoni, Sergio, Italy
 Albrecht, Ernst J., Federal
 Republic of Germany
 Albright, Hugh N., U.S.A.
 Alder, Henry L., U.S.A.
 Al-Dhahir, Mohammed W.,
 Kuwait
 Aldous, David J., U.S.A.
 Ale, Samson O., Nigeria
 Alemu, Yismaw, Ethiopia
 Alex, Leo J., U.S.A.
 Alexanderson, Gerald L.,
 U.S.A.
 Alexandrov, A. B., U.S.S.R.
 Alhussaini, Ata N., Canada
 Ali, M. Kursheed, U.S.A.
 Aljadeff, Eli, Israel
 Al-Khafaji, Mahmoud, U.S.A.
 Alladi, Krishnaswami, India
 Allday, Christopher J., U.S.A.
 Allen, Arnold O., U.S.A.
 Allen, Richard, U.S.A.
 Allotey, Francis K., Ghana
 Almeida, Jorge M. M. G.,
 Portugal
 Alperin, Jonathan L., U.S.A.
 Alperin, Roger C., U.S.A.
 Al-Rashed, Abdallah M.,
 Saudi Arabia
 Al-Salam, Waleed A.,
 Canada
 Al-Shakhs, Adnan Abdul-
 redha, Kingdom of Saudi
 Arabia
 Alsina, Claudi, Spain
 Alspach, Dale E., U.S.A.
 Alt, H. W., Federal Republic
 of Germany
 Altman, Allen B., Venezuela
 Al-Thukair, Fawzi A., Saudi
 Arabia
 Altschuler, Steven J., U.S.A.
 Alvarez, Josefina, U.S.A.
 Al-Zaid, Hassan, Kuwait
 Alzamel, Ali Iesa A., Kuwait
 Amemiya, Ichiro, Japan
 Aminfar, Habib, Iran
 Amir, Dan, Israel
 Amitsur, Shimshon A., Israel
 Anand, Bhupinder Singh, India
 Anantharaman, Claire, France
 Ancel, Fredric D., U.S.A.
 Ancona, Vincenzo, Italy
 Anderson, Douglas R., U.S.A.
 Anderson, Joel H., U.S.A.
 Anderson, John T., U.S.A.
 Anderson, Michael T., U.S.A.
 Anderson, Norman, England
 Anderson, Richard D., U.S.A.
 Anderson, Robert M., U.S.A.
 Anderssen, Robert S.,
 Australia
 Andersson, Rolf L. M., Sweden
 Andler, Martin, France
 Ando, Shiro, Japan
 Ando, Tetsuya, Japan
 Ando, Tsuyoshi, Japan
 Andrews, H. Robert, U.S.A.
 Andrews, John W., U.S.A.
 Andrié, Manfred, Federal
 Republic of Germany
 Angell, Thomas S., U.S.A.
 Anshel-Goldfeld, Iris L., U.S.A.
 Aoki, Kenji, Japan
 Aomoto, Kazuhiko, Japan
 Apostol, Tom M., U.S.A.
 Appel, Kenneth I., U.S.A.
 Appell, Jürgen, Federal
 Republic of Germany
 Apter, Arthur W., U.S.A.
 Arak, T., U.S.S.R.
 Araki, Huzihiro, Japan
 Arazy, Jonathan, Israel
 Arbarello, Enrico, Italy
 Archbold, Robert J., Scotland
 Arens, Richard F., U.S.A.
 Arika, Susumu, Japan
 Arima, Satoshi, Japan
 Arkin, Joseph, U.S.A.
 Arkowitz, Martin, U.S.A.
 Armentrout, Steve, U.S.A.
 Armitage, David H.,
 Northern Ireland
 Arnberg, Robert L., U.S.A.
 Arnold, Robert J., U.S.A.
 Arocha, Jorge L., Cuba
 Arpaia, Pasquale J., U.S.A.
 Arrhenius-Wold, Anna-Lisa,
 Sweden
 Arsenovic, Milos, Yugoslavia
 Arveson, William B., U.S.A.
 Arya, Shashi P., India
 Asam, Lilian L. C., Federal
 Republic of Germany

- Ash, Avner, U.S.A.
 Ash, J. Marshall, U.S.A.
 Ashby, Florence H., U.S.A.
 Ashour, Attia A., Egypt
 Askey, Richard A., U.S.A.
 Aslaksen, Helmer, U.S.A.
 Atiyah, Michael F., England
 Atkinson, Bruce W., U.S.A.
 Auchmuty, Giles, U.S.A.
 August, John E., U.S.A.
 Aulaskari, Rauno P., Finland
 Auslander, Bernice L., U.S.A.
 Auslander, Maurice, U.S.A.
 Avila, Geraldo S. S., Brazil
 Aviles, Patricio U., U.S.A.
 Awakeem, Nouria A., Iraq
 Axelsson, Reynir, Iceland
 Axler, Sheldon, U.S.A.
 Ayoub, Ghazi, Switzerland
 Baas, Nils A., Norway
 Baayen, Pieter C., Netherlands
 Babcock, William W., U.S.A.
 Bachar Jr., John M., U.S.A.
 Bachelis, Gregory F., U.S.A.
 Bachmann, Franz, Switzerland
 Bachmuth, Seymour, U.S.A.
 Back, Allen H., U.S.A.
 Bacopoulos, Alex, Greece
 Badamchizadeh, Alinaghi, Iran
 Bade, William G., U.S.A.
 Badger, Lee, U.S.A.
 Bagaria, Juan, U.S.A.
 Bagby, S. Thomas, U.S.A.
 Bagher Daii, Aliasghar, Iran
 Baica, Malvina F., U.S.A.
 Bailey, R. A., England
 Baillon, Jean-Bernard, France
 Baines, Michael J., England
 Bak, Anthony, Federal
 Republic of Germany
 Bakbu, Viorel, Romania
 Baker, Deidre, U.S.A.
 Baker, Kirby A., U.S.A.
 Baker, Peter F., U.S.A.
 Bakhvalov, Nikolas S., U.S.S.R.
 Balaaban, Tadeusz A., U.S.A.
 Baldin, Yuriiko Y., Brazil
 Baldinger, Robert, U.S.A.
 Ball, John M., Scotland
 Ball, Roderick D., U.S.A.
 Ballard, David J., U.S.A.
 Ballew, David W., U.S.A.
 Ballmann, Werner, U.S.A.
 Ballve, M. Eulalia, Spain
 Baloglou, George, U.S.A.
 Bamba, Siaka K., Ivory Coast
 Bamon, Rodrigo, Chile
 Banchoff, Thomas F., U.S.A.
 Bandelow, Christoph, Federal
 Republic of Germany
 Baouendi, M. Salah, U.S.A.
 Barakat, Wissam, Lebanon
 Barbeau, Edward J., Canada
 Barbosa, Joao Lucas M., Brazil
 Barel, Zeev, U.S.A.
 Barge, Marcy M., U.S.A.
 Barker, William H., U.S.A.
 Barlow, Martin T., England
 Barnes, Frank, Kenya
 Barr, Thomas H., U.S.A.
 Barrett, David E., U.S.A.
 Barrett, Lida K., U.S.A.
 Barrientos, John F., U.S.A.
 Barsky, Daniel, France
 Barth, Karl F., U.S.A.
 Barth, Theodore J., U.S.A.
 Bartle, Robert G., U.S.A.
 Barton, Susan M., U.S.A.
 Bartosiewicz, Zbigniew, Poland
 Bartoszynski, Tomek, Poland
 Baruch Jr., Herbert M., U.S.A.
 Baruch, H. Mark, U.S.A.
 Bashmakova, Izabella Grigorievna, U.S.S.R.
 Bass, Hyman, U.S.A.
 Bassily, Nader L., Egypt
 Bateman, Felice D., U.S.A.
 Bateman, Paul T., U.S.A.
 Batterson, Steve, U.S.A.
 Batty, Charles J. K., England
 Bauer, Heinz, Federal
 Republic of Germany
 Baughman, David L., U.S.A.
 Baum, Leonard E., U.S.A.
 Baumslag, Gilbert, U.S.A.
 Bayat, Gholamreza, Iran
 Bear, H. S., U.S.A.
 Beaudoin, Robert E., U.S.A.
 Beauville, Arnaud H., France
 Bebbe Njoh, Etienne,
 Cameroon
 Beck, József, Hungary
 Becker, Eberhard, Federal
 Republic of Germany
 Becker, Helmut, Federal
 Republic of Germany
 Becker, Howard S., U.S.A.
 Becker, Ronald I., South Africa
 Beechler, Barbara J., U.S.A.
 Beekmann, Wolfgang, Federal
 Republic of Germany
 Beem, John K., U.S.A.
 Beer, Gerald, U.S.A.
 Beesley, E. Maurice, U.S.A.
 Behforooz, G. H., Iran
 Behr, Kathleen, U.S.A.
 Behrends, Ehrhard, Federal
 Republic of Germany
 Behrens, Elke, Federal
 Republic of Germany
 Bekes, Robert A., U.S.A.
 Bekken, Otto B., Norway
 Bell, Howard E., Canada
 Bellenger, Joseph C., U.S.A.
 Bellile, Elke G., U.S.A.
 Belyi, G. V., U.S.S.R.
 Benardete, Diego M., U.S.A.
 Bendat, Julius S., U.S.A.
 Bennett, Donald E., U.S.A.
 Benninger, Daniel, Switzerland
 Benson, David J., U.S.A.
 Bentley, H. L., U.S.A.
 Beres, Pierre, France
 Beresina, Lyuba J., Israel
 Berezina, Mirjam T., Israel
 Berg, Christian, Denmark
 Berg, Joseph, U.S.S.R.
 Berger, Marcel, France
 Berger, Mel S., U.S.A.
 Berger, Ruth I., U.S.A.
 Berger, Thomas R., U.S.A.
 Bergman, George M., U.S.A.
 Bergqvist, Göran, Sweden
 Bergum, Gerald E., U.S.A.
 Berlekamp, Elwyn R., U.S.A.
 Berman, Lawrence, U.S.A.
 Berman, Stephen, Canada
 Bernau, Simon J., U.S.A.
 Berndt, Bruce C., U.S.A.
 Bernstein, Joseph N., U.S.A.
 Berquier, Francoise, France
 Berrick, A. J., England
 Bertaud, Marcel, Canada
 Berthelot, Pierre R., France
 Bertin, Marie José R., France
 Bertram, Aaron J., U.S.A.
 Beschler, Edwin F., U.S.A.
 Beslagic, Amer, U.S.A.
 Bessenrodt-Timmerscheidt, C.,
 Federal Republic of
 Germany
 Bessman, Marcelle, U.S.A.
 Besson, Gerard, U.S.A.
 Best, Lloyd E., U.S.A.
 Bestvina, Mladen, U.S.A.
 Beyl, F. Rudolf, U.S.A.
 Beynon, Gerard V.,
 Netherlands
 Bezdek, Andras, U.S.A.
 Bezuidenhout, Carol E., U.S.A.
 Bhatia, Rajendra, India
 Bhatnagar, Nirdosh, U.S.A.
 Biagioli, Anthony J. F., U.S.A.
 Bickel, Peter J., U.S.A.
 Bielawski, Jan P., U.S.A.
 Bielek-Schulz, Gisela, Federal
 Republic of Germany
 Bien, Frédéric V., U.S.A.
 Bierstedt, Klaus D., Federal
 Republic of Germany
 Billik, Martin, U.S.A.

- Bingen, Franz M. J., Belgium
 Bion-Nadal, Jocelyne, France
 Birch, Bryan J., England
 Birman, Joan S., U.S.A.
 Birnbaum, Z. William, U.S.A.
 Birnir, Björn, U.S.A.
 Bishop, Christopher J., U.S.A.
 Bishop, Edward E., U.S.A.
 Bishop, Wayne W., U.S.A.
 Bismut, Jean-Michel, France
 Bisson, Terrence P., U.S.A.
 Bix, Michael C., U.S.A.
 Björck, Göran, Sweden
 Björner, Anders, Sweden
 Blackadar, Bruce E., U.S.A.
 Blackwell, David, U.S.A.
 Blais, J. A. R., Canada
 Blanchard, Paul R., U.S.A.
 Blancheton, Eliane, France
 Blanton, John D., U.S.A.
 Blasberg, Steven E., U.S.A.
 Blatter, Christian, Switzerland
 Blattner, Robert J., U.S.A.
 Bleiler, Steven A., Canada
 Bloch, Ethan D., U.S.A.
 Block, Richard E., U.S.A.
 Blockus, Marilyn J., U.S.A.
 Bloom, Frederick, U.S.A.
 Blum, Lenore C., U.S.A.
 Blum, Manuel, U.S.A.
 Boardman, John M., U.S.A.
 Boas, Harold P., U.S.A.
 Bodnarescu, M. V., Federal
 Republic of Germany
 Boe, Brian D., U.S.A.
 Boehme, Thomas K., U.S.A.
 Boffi, Vinicio, Italy
 Bogart, Kenneth P., U.S.A.
 Bohn, S. Elwood, U.S.A.
 Bojarski, Bogdan, Poland
 Bokowski, Jorgen, Federal
 Republic of Germany
 Bompert, Bill E., U.S.A.
 Bonadies, Manuelita, Italy
 Bonahon, Francis, France
 Bonet, Maria Luisa, U.S.A.
 Bonnice, William E., U.S.A.
 Book, Ronald V., U.S.A.
 Booker, To Hoy, U.S.A.
 Boppa, Ravi B., U.S.A.
 Bor, Gil, U.S.A.
 Börgers, Christoph, U.S.A.
 Borho, Walter, Federal
 Republic of Germany
 Boron, Leo F., U.S.A.
 Borovoi, M. V., U.S.S.R.
 Borrelli, Robert L., U.S.A.
 Börsch-Supan, Wolfgang,
 Federal Republic of Germany
 Borwein, David, Canada
 Borwein, Peter B., Canada
 Bos, Henk J. M., Netherlands
 Bossert, J. Michael, U.S.A.
 Botelho, Fernanda, U.S.A.
 Bourgain, Jean, France
 Bourguignon, Jean-Pierre,
 France
 Bouscaren, Elisabeth, U.S.A.
 Bouvier, Alain J., France
 Bowman, Spencer B., U.S.A.
 Boyd, David W., Canada
 Boyer, Steven P., Canada
 Braden, Lawrence S., U.S.A.
 Bradford, James C., U.S.A.
 Bradlow, Steven B., U.S.A.
 Brahana, Thomas R., U.S.A.
 Brandal, Willy, U.S.A.
 Brandt, Achi, Israel
 Brandt, Rolf D., Federal
 Republic of Germany
 de Branges, Louis, U.S.A.
 Branner, Bodil, Denmark
 Branton, Richard, U.S.A.
 Breen, Stephen A., U.S.A.
 Brenti, Francesco, U.S.A.
 Breslow, Jack W., U.S.A.
 Brevit, John H., U.S.A.
 Brezzi, Franco E., Italy
 Brick, Stephen G., U.S.A.
 Briem, Eggert, Iceland
 Brieskorn, Egbert V., Federal
 Republic of Germany
 Brillhart, John D., U.S.A.
 Brin, Michael, U.S.A.
 Brink, Allen L., U.S.A.
 Brink, Daryl M., U.S.A.
 Brittenham, Mark W., U.S.A.
 Brode, John, U.S.A.
 van den Brom, Lourens,
 Netherlands
 Bronstein, Manuel, U.S.A.
 Brooks, Robert W., U.S.A.
 Brothers, John E., U.S.A.
 Broué, Michel J., France
 Broussard, Kevin, U.S.A.
 Browder, Felix, U.S.A.
 Browder, William, U.S.A.
 Brown, David W., U.S.A.
 Brown, Douglas S., U.S.A.
 Brown, Edgar H., U.S.A.
 Brown, Leon, U.S.A.
 Brown, Morton, U.S.A.
 Brown, Richard A., U.S.A.
 Brown, Robert F., U.S.A.
 Brown, Ron, U.S.A.
 Brown, Ronald, Wales
 Brownawell, W. Dale, U.S.A.
 Brownell, Frank H., U.S.A.
 Brualdi, Richard A., U.S.A.
 Bruck, Ronald E., U.S.A.
 Brulois, Frédéric P., U.S.A.
 Brumer, Armand, U.S.A.
 Brundu, Michela, Italy
 Brunk, Hugh D., U.S.A.
 Brunn, John, U.S.A.
 Brunson, Barry W., U.S.A.
 Brunswick, Natascha A.,
 U.S.A.
 Brunzie, Marion W., U.S.A.
 Bryant, Robert L., U.S.A.
 Brzezinski, Julius, Sweden
 Buchanan, Thomas, Federal
 Republic of Germany
 Buchner, Michael A., U.S.A.
 Buckley, Joseph, U.S.A.
 Buechler, Steven A., U.S.A.
 Bueker, Robert C., U.S.A.
 Buhler, Joe P., U.S.A.
 Buianoukas, Francis R.,
 U.S.A.
 Bujalance, Emilio, Spain
 Bulena, Richard O., U.S.A.
 Bull, Everett L., U.S.A.
 Bumby, Richard T., U.S.A.
 Burak, Tamar, Israel
 Burbridge, Thomas N., U.S.A.
 Burckel, Robert B., U.S.A.
 Burgess, C. Edmund, U.S.A.
 Burgstahler, Sylvan, U.S.A.
 Burgués, Carmen, Spain
 Burke, Douglas R., U.S.A.
 Burke, John R., U.S.A.
 Burkhart, Richard H., U.S.A.
 Burman, Prabir, U.S.A.
 Burns, Robert G., Canada
 Burr, Stefan A., U.S.A.
 Burton, T. A., U.S.A.
 Buss, Samuel R., U.S.A.
 Butler, Jeffrey, U.S.A.
 Butler, Loren, U.S.A.
 Butz, Jeffrey R., U.S.A.
 Buyske, Steven G., U.S.A.
 Byers, Bill, Canada
 Byrkit, Donald R., U.S.A.
 Cable, Charles A., U.S.A.
 Cagnac, Francis, Cameroon
 Calabi, Eugenio, U.S.A.
 Caldwell, Chris K., U.S.A.
 Caldwell, James, England
 Calhoun, William C., U.S.A.
 Call, Gregory S., U.S.A.
 Callahan, Susan L., U.S.A.
 Calvert, Bruce D., New Zealand
 Calvis, David T., U.S.A.
 Cambern, Michael J., U.S.A.
 Campbell, Colin M., Scotland
 Campbell, Eddy, Canada
 Campbell, Neville, U.S.A.
 Campbell, Paul J., U.S.A.
 Campbell, Robert I., U.S.A.

- Cancelier, Claudy, West Indies
 Cannon, James W., U.S.A.
 Cannon, John J., Australia
 Cantwell, John C., U.S.A.
 Canuto, Claudio, Italy
 Cao, Huai-Dong, People's Republic of China
 Cappell, Sylvain E., U.S.A.
 Cargo, Gerald T., U.S.A.
 Carlson, James A., U.S.A.
 Carlson, Robert C., U.S.A.
 Carlsson, Gunnar E., U.S.A.
 Carlsson, Renate, Federal Republic of Germany
 Carlton, Eloise H., U.S.A.
 Carmichael, Jean-Pierre, Canada
 Carney, Rose A., U.S.A.
 Carnielli, Walter A., Brazil
 Carr, James G., U.S.A.
 Carrier, Stephen, U.S.A.
 Carroll, Catherine R., U.S.A.
 Carter, Jack A., U.S.A.
 Carter, Roger W., England
 Cartier, Pierre E., France
 Carton, Edmond J., Belgium
 Carton-Lebrun, C., Belgium
 Case, Bettye A., U.S.A.
 Casgrain, Louis, Canada
 Cassis, Beatriz, U.S.A.
 Casson, Andrew J., U.S.A.
 van Casteren, Johannes A., Belgium
 Castro, Helena M. A., Brazil
 Catlin, David W., U.S.A.
 Cattani, Eduardo H., U.S.A.
 Cavaretta, Alfred S., U.S.A.
 Cawley, Elise E., U.S.A.
 Cayford, Afton H., Canada
 Ceccherini, Pier Vittorio, Italy
 Cecil, Thomas E., U.S.A.
 Cendra, Hernan, U.S.A.
 Cercignani, Carlo, Italy
 Chacron, Maurice, Canada
 Chademan, Arsalan, Iran
 Chahal, Jasbir S., U.S.A.
 Chambon, Claude, France
 Chan, Jean B., U.S.A.
 Chan, Wai-Kwan, Hong Kong
 Chandra, Jagdish, U.S.A.
 Chang, Der-chen E., U.S.A.
 Chang, Jen-Tseh, U.S.A.
 Chang, Kun Soo, South Korea
 Chang, Shao-Chien, Canada
 Chang, Sheldon X., U.S.A.
 Chang, Sun-Yung Alice, U.S.A.
 Channell, Paul J., U.S.A.
 Chao, Jia-Arng, U.S.A.
 Chapline, George F., U.S.A.
 Charney, Ruth M., U.S.A.
 Chatterjee, Gautam, U.S.A.
 Chatterji, Shristi D., Switzerland
 Chattopadhyay, Rita, U.S.A.
 Chatzidakis, Zoé M., U.S.A.
 Chazin, Robert L., U.S.A.
 Cheeger, Jeff, U.S.A.
 Chein, Orin N., U.S.A.
 Chen, Daniel K. C., U.S.A.
 Chen, Haiwen, U.S.A.
 Chen, Li-Chen, U.S.A.
 Chen, Lung-Kee, U.S.A.
 Chen, William W.L., England
 Chen, Young-Ming, U.S.A.
 Cheney, Margaret, U.S.A.
 Cheng, Katherine H., U.S.A.
 Cheng, Kuo-Shung, China-Taiwan
 Cheng Min De, People's Republic of China
 Cherlin, Gregory L., U.S.A.
 Chern, Shiing Shen, U.S.A.
 Chernoff, Paul R., U.S.A.
 Cheung, Chi-Keung, U.S.A.
 Cheung, Samson Hokchi, U.S.A.
 Chevallier, Benoît, France
 Chi, Quo-Shin, U.S.A.
 Chichester, Frederick D., U.S.A.
 Chien Mao-Ting, China-Taiwan
 Childs, Lindsay N., U.S.A.
 Chin, Watson L., U.S.A.
 Chinn, William G., U.S.A.
 Cho, Tae-Geun, South Korea
 Choe, Geon Ho, U.S.A.
 Choie, Youngju, U.S.A.
 Chong Chi-Tat, Singapore
 Chorin, Alexandre J., U.S.A.
 Chou Ching-Sung, China-Taiwan
 Chou, Shang-Ching, U.S.A.
 Chow, Bennett, U.S.A.
 Chow, Pao-Liu, U.S.A.
 Chowdhury, Munibur R., Bangladesh
 Christ, Lily E., U.S.A.
 Christ, Mike, U.S.A.
 Christy, Joseph P., U.S.A.
 Chu, C. H., England
 Chuang, Chen-Lian, U.S.A.
 Chun, Chong Suh, Canada
 Chung, Fan R. K., U.S.A.
 Chung, Lung O., U.S.A.
 Church, Philip T., U.S.A.
 Ciesielski, Zbigniew, Poland
 Ciliberto, Carlo, Italy
 Cima, Joseph, U.S.A.
 Ciment, Melvyn, U.S.A.
 Cinlar, Erhan, U.S.A.
 Cipra, Barry A., U.S.A.
 Clapp, Monica, Mexico
 Clark, Jack, U.S.A.
 Clark, John, New Zealand
 Clark, Judith A., U.S.A.
 Clarke, A. Bruce, U.S.A.
 Clarke, Wilton E. L., U.S.A.
 Clay, Oliver K., Switzerland
 Clemens, Herbert, U.S.A.
 Clemons, Arthur J., U.S.A.
 Clermont, Michèle, Canada
 Clover, William, U.S.A.
 Clozel, Laurent, France
 Cochran, Tim D., U.S.A.
 Coghlan, Leslie, U.S.A.
 Cohen, D. I. A., U.S.A.
 Cohen, Frederick R., U.S.A.
 Cohen, Moses E., U.S.A.
 Cohen, Paul J., U.S.A.
 Cohen, Ralph L., U.S.A.
 Cohn, Harvey, U.S.A.
 Cohn, Leslie, U.S.A.
 Cohn, Paul M., England
 Cohn, Richard M., U.S.A.
 Cohoon, David K., U.S.A.
 Cole, Nancy, U.S.A.
 Coleman, Courtney S., U.S.A.
 Coleman, Robert F., U.S.A.
 Colin de Verdière, Yves, France
 Colley, Susan J., U.S.A.
 Collins, George E., U.S.A.
 Collins, Peter J., England
 Colliot-Thélène, Jean-Louis, France
 Colón, Nilsa, U.S.A.
 Colvin, Michael R., U.S.A.
 Combes, Francois, France
 Comer, Stephen D., U.S.A.
 Comerford, Leo P., U.S.A.
 Comfort, W. W., U.S.A.
 Concus, Paul, U.S.A.
 Conlan, James, Canada
 Conlon, Lawrence W., U.S.A.
 Connell, Edwin H., U.S.A.
 Connelly, Robert, U.S.A.
 Conrad, Bruce P., U.S.A.
 Contessa, Maria, Italy
 Cooke, Kenneth L., U.S.A.
 Cooney, Miriam P., U.S.A.
 Cooper, Daryl, U.S.A.
 Cooper, Duane A., U.S.A.
 Cooperman, Gene D., U.S.A.
 Cootz, Thomas A., U.S.A.
 Corach, Gustavo, Argentina
 Corbett, Dorothy S., U.S.A.
 Corcoran, John, U.S.A.
 Cordes, Eva R., U.S.A.

- Cordes, H. O., U.S.A.
 Corduneanu, Constantin C., U.S.A.
 Cornea, Aurel, Federal Republic of Germany
 Cornell, Gary, U.S.A.
 Cornet, Bernard, France
 Corzatt, Clifton E., U.S.A.
 Coste, Michel, France
 Costenoble, Steven R., U.S.A.
 Coste-Roy, Marie-Francoise, France
 Côté, Normand H., U.S.A.
 Cotner, Carl F., U.S.A.
 Coutsiass, Evangelos A., U.S.A.
 Coutu, Gerard, U.S.A.
 Cowen, Carl C., U.S.A.
 Cowens, J. Wayne, U.S.A.
 Cox, Milton D., U.S.A.
 Cozzens, John H., U.S.A.
 Crabtree, James B., U.S.A.
 Craig, Walter, U.S.A.
 Craig, William, U.S.A.
 Crapo, Henry, France
 Cree, George C., Canada
 Creighton, James H. C., U.S.A.
 Crilly, Anthony J., Hong Kong
 Croke, Christopher B., U.S.A.
 Croom, Frederick H., U.S.A.
 Cudia, Dennis F., U.S.A.
 Cull, Paul, U.S.A.
 Cullen, Robert, U.S.A.
 Culler, Marc, U.S.A.
 Culverhouse, Robert C., U.S.A.
 Currey, Bradley N., U.S.A.
 Curtis, Charles W., U.S.A.
 Curtis, Philip C., U.S.A.
 Cusick, Larry W., U.S.A.
 Cutler, Tom, U.S.A.
 Cuyt, Annie A.M., Belgium
 Daboussi, Hedi, France
 Dad-Del, Aliakbar, U.S.A.
 Dadok, Jiri, U.S.A.
 Dahiya, R. S., U.S.A.
 Dahlquist, Germund G., Sweden
 Dai, Xianzhe, U.S.A.
 Dalang, Robert C., Switzerland
 Dales, Harold G., England
 D'Ambrosio, Ubiratan, Brazil
 Damiano, David B., U.S.A.
 Damon, James N., U.S.A.
 Dancis, Jerome, U.S.A.
 Dankel, Thad G., U.S.A.
 D'Antini, Umberto M., Canada
 Dargahi-Noubary, Gholam-reza, Iran
 Darken, Joanne S., U.S.A.
 Darmon, Henri R., Canada
 Dastrange, Nasser, Iran
 Daughtry, John, U.S.A.
 Davarashvili, Margaret, Israel
 Daverman, Robert J., U.S.A.
 David, Guy R., France
 Davidon, William C., U.S.A.
 Davidson, Kenneth R., Canada
 Davie, Alexander M., Scotland
 Davis, Adrienne, U.S.A.
 Davis, B. Mark, U.S.A.
 Davis, Chandler, Canada
 Davis, James F., U.S.A.
 Davis, Mark H. A., England
 Davis, Martin D., U.S.A.
 Davis, Philip J., U.S.A.
 Davis, Ronald M., U.S.A.
 Davison, J. Leslie, Canada
 Dawson, John W., U.S.A.
 Dawson, Michael J., U.S.A.
 Day, Jane M., U.S.A.
 Dayton, Barry H., U.S.A.
 De Caen, D., Canada
 De Concini, Corrado, Italy
 De Graaf, Jan, Netherlands
 De Lange, Jan, Netherlands
 De Souza, Paulo Ney R., U.S.A.
 De Wilde, Philippe, Belgium
 Debnath, Lokenath, U.S.A.
 Dechamps, Myriam, France
 Dechéne, Lucy I., U.S.A.
 Decker, Wolfram, U.S.A.
 Deeb, Waleed M., Kuwait
 Dehnad, Khosrow, U.S.A.
 Dekker, Jacob C. E., U.S.A.
 Dekster, Boris V., Canada
 Del Riego, Lilia, Mexico
 Delaney, Matthew S., U.S.A.
 Delaware, Richard, U.S.A.
 Delzell, Charles N., U.S.A.
 Demana, Franklin D., U.S.A.
 Deng, Bo, U.S.A.
 Dennis, R. Keith, U.S.A.
 Depue, Bill E., U.S.A.
 Deretsky, Tatiana, U.S.A.
 Derrig, Richard A., U.S.A.
 Deschamps, Claude, France
 Desmedt, Yvo G., Belgium
 Desquith, Etienne, Ivory Coast
 Deumens, Erik, U.S.A.
 Devaney, Robert L., U.S.A.
 Deveney, James K., U.S.A.
 Diaconis, Persi W., U.S.A.
 Dias-Agudo, Fernando R., Portugal
 Dickenstein, Alicia M., Argentina
 Dieck, Tammo, Federal Republic of Germany
 Dieu, Phan Dinh, Vietnam
 Dilcher, Karl, Canada
 Diliberto, Stephen P. L., U.S.A.
 Dimitrić, Radoslav M., Yugoslavia
 Dimovski, Dončo, Yugoslavia
 Dinculeanu, Nicolae, U.S.A.
 Ding Er-sheng, People's Republic of China
 Ding Ren, People's Republic of China
 Dinh, Hung T., U.S.A.
 Dinneen, Gerald P., U.S.A.
 Dionne, Philippe-Auguste, France
 DiPerna, Ronald J., U.S.A.
 Dixon, John D., Canada
 Djoković, Dragomir Ž., Canada
 Djomehri, M. Jahed, U.S.A.
 Dlab, Vlastimil, Canada
 Dobber, Eelkje, Netherlands
 Dobbins, Robert R., U.S.A.
 Dodson, Christopher T. J., England
 Dokmeci, M. Cengiz, Turkey
 Dolgachev, Igor, U.S.A.
 Dollard, John D., U.S.A.
 Dollinger, Michael R., U.S.A.
 Donaldson, James A., U.S.A.
 Donaldson, Simon K., England
 Donelan, Peter S., New Zealand
 Doner, John E., U.S.A.
 Donnay, Victor J., U.S.A.
 Donnellan, John R., U.S.A.
 Donoghue, William F., U.S.A.
 Donoho, David L., U.S.A.
 Doran, Robert S., U.S.A.
 Dorman, David R., U.S.A.
 Dotseth, Gregory M., U.S.A.
 Doughty, Herb, U.S.A.
 Douglas, Robert J., U.S.A.
 Douglas, Ronald G., U.S.A.
 Doyle, John C., U.S.A.
 Doyle, Peter G., U.S.A.
 Drábek, Pavel, Czechoslovakia
 Draper, Patricia, U.S.A.
 Draper, Richard N., U.S.A.
 Dreiling, Leslie A., U.S.A.
 Driessell, Kenneth R., U.S.A.
 Drinfel'd, V. G., U.S.S.R.
 Dristy, Forrest E., U.S.A.
 Driver, Bruce K., U.S.A.
 Drobnies, Saul I., U.S.A.
 Drobot, Vladimir, U.S.A.
 Drucker, Thomas L., U.S.A.
 D'Souza, Harry J., U.S.A.
 Dubejko, Maciej, Nigeria
 Dubinsky, Ed, U.S.A.
 Dubins, Lester E., U.S.A.
 Dubois, Jacques, Canada

- Dudley, Underwood, U.S.A.
 Duff, George F. D., Canada
 Duke, Carla R., U.S.A.
 Dulin, Bill J., U.S.A.
 Dummit, David S., U.S.A.
 Dunham, Douglas J., U.S.A.
 Dunkl, Charles F., U.S.A.
 Dunninger, Dennis R., U.S.A.
 Dunwoody, James, Northern
 Ireland
 DuRapau, Victor J., U.S.A.
 Durfee, Alan H., U.S.A.
 Durst, Lincoln, U.S.A.
 Dutko, Michael, U.S.A.
 Duyos, Sarah M., Cuba
 Dydak, Jerzy, U.S.A.
 Dyer, Matthew J., Australia
 Dymacek, Wayne M., U.S.A.
 Dynin, Alexander, U.S.A.
 Earnest, Andrew G., U.S.A.
 Eckhoff, Jürgen, Federal
 Republic of Germany
 Eckmann, Beno, Switzerland
 Eckmann, Jean-Pierre,
 Switzerland
 Edelson, Allan L., U.S.A.
 Edgar, Gerald A., U.S.A.
 Edgar, Hugh M., U.S.A.
 Edward, Harold M., U.S.A.
 Edwards, Robert D., U.S.A.
 Ee, Soo-Mei, Malaysia
 Eells, James, England
 Eenigenburg, Paul J., U.S.A.
 Effros, Edward G., U.S.A.
 Efrat, Isaac Y., U.S.A.
 Egghe, Leo C., Belgium
 Eggleston, Trent, U.S.A.
 Egoryčev, Georgii P., U.S.S.R.
 Eguchi, Kazuo, Japan
 Ehrlich, Paul E., U.S.A.
 Ehrlich, Stefan, U.S.A.
 Ein, Lawrence M. H., U.S.A.
 Einwohner, Theodore H.,
 U.S.A.
 Eirola, Timo Juhani, Finland
 Eisele, Carolyn, U.S.A.
 Eklof, Paul C., U.S.A.
 Elderkin, Richard H., U.S.A.
 Eliashberg, Ya. M., U.S.S.R.
 Elliott, David, Australia
 Elliott, George A., Denmark
 Ellis, Alan J., Hong Kong
 Ellis, Graham J., England
 Ellis Jr., Wade, U.S.A.
 Elwin, John D., U.S.A.
 Embid, Pedro F., U.S.A.
 Emsalem, Michel, France
 Enderton, Herbert B., U.S.A.
 Endo, Youichi, U.S.A.
 van Engelen, Fons, Netherlands
 Engler, Antonio, Brazil
 Engler, Hans P., U.S.A.
 Enguehard, Michel E., France
 Enock, Michel, France
 Enomoto, Kazuyuki, Japan
 Epelbaum, Yonathan, U.S.A.
 Epstein, David B. A.,
 England
 Epstein, Mordechai, Israel
 Epstein, Raisa, Israel
 Erbland, John P., U.S.A.
 Erdoğan, Vahap, Bahrain
 Erdős, Paul, Hungary
 Erez, Boas, Switzerland
 Eriksson, Folke S., Sweden
 Eriksson, Sirkka-Liisa, Finland
 Erkama, Timo, Finland
 Erlandsson, Thomas E. V.,
 Sweden
 Eryvnyck, Gontran J., Belgium
 Eschenbach, Dieter, Federal
 Republic of Germany
 Eskin, Gregory, U.S.A.
 Esnault, Hélène I., Federal
 Republic of Germany
 Evans, Lawrence C., U.S.A.
 Evans, Michael J., Canada
 Evens, Leonard, U.S.A.
 Everett, Jeff, U.S.A.
 Ewing, John H., U.S.A.
 Ewy, Daniel J., U.S.A.
 Eyles, Joseph W., U.S.A.
 Fabrikant, Valery I., Canada
 Faddeev, Ludvig D., U.S.S.R.
 Faires, Barbara T., U.S.A.
 Faltings, Angelika, U.S.A.
 Faltings, Gerd, U.S.A.
 Fan Chun-Tak, Hong Kong
 Fan, Peng, U.S.A.
 Fang, Joong J., U.S.A.
 Faour, Nazih Sami, Kuwait
 Faran, James J., U.S.A.
 Farran, Hani R., Kuwait
 Farrand, Scott M., U.S.A.
 Farris, Frank A., U.S.A.
 Fässler, Albert F., Switzerland
 Fauci, Lisa J., U.S.A.
 Faure, Robert, France
 Favro, R. G., U.S.A.
 Faxén, Birger J., Sweden
 Fayn, Boris, U.S.A.
 Federer, Leslie J., U.S.A.
 Fedortchouk, Vitali V.,
 U.S.S.R.
 Feehan, Paul, U.S.A.
 Feferman, Solomon, U.S.A.
 Fefferman, Charles L., U.S.A.
 Feichtinger, Hans G., Austria
 Fein, Burton I., U.S.A.
 Feingold, Alex, U.S.A.
 Fekete, Alan D., U.S.A.
 Feldman, David V., U.S.A.
 Feldman, Jacob, U.S.A.
 Feldman, Leonard, U.S.A.
 Feldman, Mark, U.S.A.
 Fellin, Jo Ann, U.S.A.
 Fenn, Roger A., England
 Ferguson, Helaman R.P.,
 U.S.A.
 Feroe, John A., U.S.A.
 Ferrer-Santos, Walter R.,
 Uruguay
 Fiedler, Miroslav, Czechoslovakia
 Figá-Talamanca, Alessandro,
 Italy
 Figiel, Tadeusz, Poland
 Fillmore, Jay P., U.S.A.
 Filus, Jerzy K., U.S.A.
 Filus, Lidia Z., U.S.A.
 Finbow, Arthur S., Canada
 Fine, Benjamin, U.S.A.
 Fine, Nathan J., U.S.A.
 Finkelstein, Larry A., U.S.A.
 Finstow, David R., U.S.A.
 Fintushel, Ronald A., U.S.A.
 Fischer, Arthur E., U.S.A.
 Fischer, Gerd O. W., Federal
 Republic of Germany
 Fischer-Colbrie, Doris H.,
 U.S.A.
 Fisher, Joe W., U.S.A.
 Fisher, Newman H., U.S.A.
 Fisk, Steve T., U.S.A.
 Fjelstad, Paul, U.S.A.
 Flaherty, Francis J., U.S.A.
 Flanders, Harley, U.S.A.
 Flapan, Erica L., U.S.A.
 Flashman, Martin, U.S.A.
 Flath, Dan, Singapore
 Flatto, Leopold, U.S.A.
 Flegmann, Alex, England
 Fleischer, I., U.S.A.
 Fleming, Richard J., U.S.A.
 Flesner, David E., U.S.A.
 Fletcher, Evan M., U.S.A.
 Florenzano, Monique, France
 Flores, Consuelo, Nicaragua
 Flowers, Joe D., U.S.A.
 Folland, Gerald B., U.S.A.
 Fong, Che-Kao, Canada
 Fong, Paul, U.S.A.
 Fontaine, Jean-Marc, France
 Fontana, Marco, Italy
 Fontes-Sousa, Fatima,
 Portugal
 Ford, Charles E., U.S.A.
 Ford, Richard L., U.S.A.
 Foreman, Matthew D., U.S.A.
 Forman, Robin, U.S.A.

Förster, Karl-Heinz, Federal

Republic of Germany

Forster, Otto F., Federal

Republic of Germany

Forsythe, W. L., U.S.A.

Foster, Andrew, U.S.A.

Foster, Lee, U.S.A.

Foster, Lorraine L., U.S.A.

Fournier, Gilles, Canada

Fournier, John J.F., Canada

Fowler, Peter A., U.S.A.

Fox, Abigail J., U.S.A.

Franco, Zachary M., U.S.A.

Franco-Sánchez, Ernesto,
U.S.A.

Frank, Marguerite J., U.S.A.

Frankl, Peter, France

Franks, John M., U.S.A.

Frayne, Thomas E., U.S.A.

Freed, Daniel S., U.S.A.

Freedman, Michael H., U.S.A.

Freidkin, Solomon, Israel

Freire, Alexandre S., U.S.A.

Freitag, Herta T., U.S.A.

Frenkel, Igor B., U.S.A.

Fridman, Buma L., U.S.A.

Friedland, Shmuel, U.S.A.

Friedlander, Eric M., U.S.A.

Friedlander, John B., Canada

Friedlander, Susan J., U.S.A.

Friedman, Eduardo C., U.S.A.

Friedman, Yaakov, U.S.A.

Frih, Elmostapha, Canada

Fröhlich, Jürg M., Switzerland

Frota-Mattos, Luiz A., Brazil

Fu, Joseph H. G., U.S.A.

Fu, Lorraine, U.S.A.

Fuglede, Bent, Denmark

Fujita, Hiroshi, Japan

Fujiwara, Takeshi, Japan

Fukui, Toshizumi, Japan

Fuller, William R., U.S.A.

Funkenbusch, Walter W.,

U.S.A.

Furness, Dewey F., U.S.A.

Furusawa, Harushi, Japan

Furuta, Mikio, Japan

Fusaro, B. A., U.S.A.

Fyfe, David J., England

Gaal, Lisl N., U.S.A.

Gabai, David, U.S.A.

Gabriel, Peter M., Switzerland

Gadbois, Steven C., U.S.A.

Gaede, Karl W., Federal

Republic of Germany

Gaffney, Matthew P., U.S.A.

Gaillard, Pierre Y., U.S.A.

Galajda, Pavel, Czechoslovakia

Galdon, Jose M., Spain

Gale, David, U.S.A.

Gallavotti, G., Italy

Galloway, Gregory J., U.S.A.

Galovich, Steven P., U.S.A.

Galperin, Efim A., Canada

Games, Richard A., U.S.A.

Gangolli, Ramesh A., U.S.A.

Ganong, Richard, Canada

Gao, Rui, People's Republic
of China

Garcia, Mariano, U.S.A.

Gardiner, Anthony, England

Gardiner, Stephen J., Ireland

Gardner, Maria F., U.S.A.

Gardner, Robert B., U.S.A.

Garfunkel, Solomon A., U.S.A.

Garnett, John B., U.S.A.

Garrison, Betty B., U.S.A.

Garzon, Max, U.S.A.

Gasper, George, U.S.A.

Gass, Michael D., U.S.A.

Gaudard, Marie, U.S.A.

Gauthier, Paul M., Canada

Gawędzki, Krzysztof, France

Gehring, Frederick W., U.S.A.

Gelbart, Stephen S., Israel

Geller, Daryl N., U.S.A.

Geller, Susan C., U.S.A.

Geller, William, U.S.A.

Geman, Stuart A., U.S.A.

Geoghegan, Ross, U.S.A.

Gerardin, Paul, U.S.A.

Gerber, Christoph, Switzerland

Gerber, P. Dean, U.S.A.

Gerlach, Eberhard, U.S.A.

Germay, Noël, Belgium

Geronimo, Jeffrey S., U.S.A.

Gerstein, Larry J., U.S.A.

Gerstenhaber, Murray, U.S.A.

Gerth, Frank E., U.S.A.

Gerver, Joseph L., U.S.A.

Getu, Seyoum, U.S.A.

Getzler, Ezra, France

Ghosh, P. R., India

Giaquinta, Mariano, Italy

Gibson, Peter M., U.S.A.

Gilat, David, Israel

Gilberg, David, U.S.A.

Gilbert, Nick D., Wales

Gildenhuis, Dion, Canada

Giles, John R., Canada

Gilfeather, Frank, U.S.A.

Gillet, Henri A., U.S.A.

Gilligan, Bruce C., Canada

Gillman, Leonard, U.S.A.

Gilman, Jane P., U.S.A.

Gilman, Robert H., U.S.A.

Gilmer, Robert, U.S.A.

Gilpin, Michael J., U.S.A.

Gimbel, John G., Denmark

Ginn, Tong, U.S.A.

Ginzburg, Victor A., U.S.S.R.

Girolo, Jack E., U.S.A.

Giusti, Marc F., France

Givens, James W., U.S.A.

Glaser, Moses, U.S.A.

Glass, Michael S., U.S.A.

Glauber, George I., U.S.A.

Gleason, Andrew M., U.S.A.

Glidden, Bridget B., U.S.A.

Gloor, Jim, U.S.A.

Glowinski, Roland, U.S.A.

Gluskin, E. D., U.S.S.R.

Goberstein, Simon M., U.S.A.

Godbole, Anant P., U.S.A.

Godounov, Sergei K., U.S.S.R.

Gokhale, Jyotsna, U.S.A.

Golbabali, Ahmold, Iran

Goldbach, Ronald W., Netherlands

Goldberg, Donald Y., U.S.A.

Goldberg, Lisa R., U.S.A.

Goldberg, Samuel, U.S.A.

Goldberg, Vladislav V., U.S.A.

Goldblatt, Robert I., New
Zealand

Goldfeld, Dorian, U.S.A.

Goldhaber, J. K., U.S.A.

Goldman, Jay R., U.S.A.

Goldman, Malcolm, U.S.A.

Goldschmidt, David M., U.S.A.

Goldstein, Martin I., Canada

Goldstein, Myron, England

Goldstein, Richard Z., U.S.A.

Goldston, Daniel A., U.S.A.

Goldwasser, Shafi, U.S.A.

Gomez-Larrañaga, Jose-Carlos,
Mexico

Gomph, Robert E., U.S.A.

Gonchar, A. A., U.S.S.R.

Gondard, Danielle J., France

Gonshor, Harry, U.S.A.

Gonzalo, Jesus, U.S.A.

Good, Richard A., U.S.A.

Goodman, A. W., U.S.A.

Goodman, F., U.S.A.

Goodman, Gerald S., Italy

Goodman, Jacob E., U.S.A.

Goodman, Richard H., U.S.A.

Goodman, Roe W., U.S.A.

Goodrick, Richard E., U.S.A.

Goodson, Geoffrey R., South

Africa

Gordon, B. Brent, U.S.A.

Gordon, Cameron M., U.S.A.

Gorenflo, Rudolf, Federal

Republic of Germany

Goroff, Daniel L., U.S.A.

Gottlieb, Daniel H., U.S.A.

Gould, Gerald G., Wales

Goyal, Vinod B., U.S.A.

Grabner, David J., U.S.A.

- Grabiner, Judith V., U.S.A.
 Grabiner, Sandy, U.S.A.
 Grace, E. E., U.S.A.
 Grace, Winston R., U.S.A.
 Graf, Siegfried, U.S.A.
 Grafakos, Loukas G., U.S.A.
 Graham, Ian R., Canada
 Graham, Ronald L., U.S.A.
 Graham, Victor W., Ireland
 Graham-Eagle, James, U.S.A.
 Granas, Andrzej, Canada
 Grant, David R., U.S.A.
 Grant, Douglass L., Canada
 Grässer, H. Siegfried P., South Africa
 Gravel, Pierre, Canada
 Graver, Jack E., U.S.A.
 Graves, G. Elton, U.S.A.
 Gray, Mary W., U.S.A.
 Grayson, Daniel R., U.S.A.
 Grayson, Matthew A., U.S.A.
 Green, Willie H., U.S.A.
 Greenfield, Gary R., U.S.A.
 Greengard, Claude A., U.S.A.
 Greenspoon, Arthur, U.S.A.
 Greguš, Michal, Czechoslovakia
 Gretskey, Neil, U.S.A.
 Griess, Robert L., U.S.A.
 Grigorev, D. Yu, U.S.S.R.
 Grillakis, Manoussos, U.S.A.
 Grinstead, Charles M., U.S.A.
 Grobler, Jacobus J., South Africa
 Gromov, Mikhael, France
 Grone, Robert D., U.S.A.
 Gross, Benedict H., U.S.A.
 Gross, Fred, U.S.A.
 Gross, Kenneth I., U.S.A.
 Gross, Louis J., U.S.A.
 Grossberg, Rami P., U.S.A.
 Grosser, Siegfried K., Austria
 Grossinho, Maria do R., Portugal
 Grossman, Jerrold W., U.S.A.
 Grosswald, Emil, U.S.A.
 Groszek, Marcia J., U.S.A.
 Grove, Karsten, U.S.A.
 Grove, Larry C., U.S.A.
 Grubb, Daniel J., U.S.A.
 Gruenig, Thompson DeP., U.S.A.
 Grunbaum, F. Alberto, U.S.A.
 Grundman, Helen G., U.S.A.
 Grüter, Michael G., Federal Republic of Germany
 Grzegorzczak, Iwona M., U.S.A.
 Grzegorzczak, Piotr, Poland
 Gu Chaohao, People's Republic of China
 Guan, Puhua, U.S.A.
 Guckenheimer, John M., U.S.A.
 Guidy Wandja, Josephine, Ivory Coast
 Guillen, Francisco, Spain
 Guiro, Abdoulaye, Senegal
 Gunawardena, Jeremy H. C., England
 Günaydin, Murat, U.S.A.
 Gundlach, Karl-Bernhard, Federal Republic of Germany
 Gunn, Charles, U.S.A.
 Gupta, Laxmi N., U.S.A.
 Gupta, Rajiv, Canada
 Gupta, Shiv K., U.S.A.
 Gurak, Stanley J., U.S.A.
 Guralnick, Robert M., U.S.A.
 Guseman, Larry F., U.S.A.
 Gutowitz, Howard, U.S.A.
 Györy, Kálmán, Hungary
 Haagerup, Uffe, Denmark
 Hada, Ichiro, Japan
 Hadid, Sameer, Iraq
 Haeberly, Jean-Pierre A., U.S.A.
 Hafner, James L., U.S.A.
 Hag, Kari J., Norway
 Hag, Per, Norway
 Hager, Harold W., U.S.A.
 Hahn, Alexander J., U.S.A.
 Hailat, Mohammad Q., Jordan
 Haiman, Mark, U.S.A.
 Haimo, Deborah T., U.S.A.
 Hajela, D.J., U.S.A.
 Hakeda, Jôsukey, Japan
 Halberstam, Heini, U.S.A.
 Hales, Alfred W., U.S.A.
 Hales, Thomas C., U.S.A.
 Hall, Glen R., U.S.A.
 Hall, Peter, U.S.A.
 Halmos, Paul R., U.S.A.
 Halpern, Herbert, U.S.A.
 Hamann, Eloise A., U.S.A.
 Hamilton, Richard, U.S.A.
 Hammond, William F., U.S.A.
 Hamrick, Gary C., U.S.A.
 Han, Sang-Geun, U.S.A.
 Hancock, Don L., U.S.A.
 Hancock, Helen M., U.S.A.
 Handel, David, U.S.A.
 Hanes, Kit, U.S.A.
 Hanges, Nicholas, U.S.A.
 Hansen, Idar, Norway
 Hansen, Vagn L., Denmark
 Harada, Masana, U.S.A.
 Harari, Sami R., France
 Harary, Frank, U.S.A.
 Harbater, David, U.S.A.
 Harborth, Heiko, Federal Republic of Germany
 Hardie, Keith A., South Africa
 Hardin, Douglas P., U.S.A.
 Hardt, Robert M., U.S.A.
 Harel, Michel, France
 Har'el, Zvi, U.S.A.
 Harkleroad, Edie, U.S.A.
 Harkleroad, Leon W., U.S.A.
 Harrington, Leo A., U.S.A.
 Harris, Bruno, U.S.A.
 Harris, David C., U.S.A.
 Harris, John C., U.S.A.
 Harris, Michael H., U.S.A.
 Harris, Steven G., U.S.A.
 Harrison, Jenny, U.S.A.
 Hart, Damon, U.S.A.
 Hart, Mikal, U.S.A.
 Hart, Neal, U.S.A.
 Hart, Sergiu, Israel
 Hartnett, William E., U.S.A.
 Harzheim, Egbert R., Federal Republic of Germany
 Hasegawa, Keizo, U.S.A.
 Hashemi-Parast, Sayed M., Iran
 Haslinger, Friedrich, U.S.A.
 Hastaa, Johan, U.S.A.
 Hastings, Maryam S., U.S.A.
 Hatcher, Theodore R., U.S.A.
 Hatcher, William S., Canada
 Hatori, Asako, Japan
 Hatori, Osamu, Japan
 Hatt-Arnold, Didier, Switzerland
 Hausmann, Jean-Claude R., Switzerland
 Haussermann, John W., U.S.A.
 Haussmann, Werner, Federal Republic of Germany
 Hawkins, Betty L., U.S.A.
 Hawkins, John H., U.S.A.
 Hawkins, Thomas W., U.S.A.
 Hawkins, William A., U.S.A.
 Hawn, Virginia S., U.S.A.
 Hayashi, Eric, U.S.A.
 Hayashi, Hiroshi, Japan
 Hayes, David F., U.S.A.
 Hays, Dorothy A., U.S.A.
 Hazewinkel, Michiel, Netherlands
 Head, Richard R., U.S.A.
 Heath, Steven H., U.S.A.
 Heath-Brown, David R., England
 Hedberg, Lars Inge, Sweden
 Heiligman, Mark I., U.S.A.
 Heinzer, William J., U.S.A.
 Heitsch, James L., U.S.A.
 Hejhal, Dennis A., U.S.A.
 Hekster, Nick, Netherlands
 Helgason, Sigurdur, U.S.A.
 Helton, Bill, U.S.A.
 Helton, F. Joanne, U.S.A.

- Helversen-Pasotto, Anna,
 France
 Henderson, Francis McV.,
 U.S.A.
 Henderson, James P., U.S.A.
 Hendon, Mo D., U.S.A.
 Henkin, Leon, U.S.A.
 Henrichs, Rold W., Federal
 Republic of Germany
 Henry, Jean-Pierre G., France
 Hensley, Douglas A., U.S.A.
 Hensley, Elmer L., U.S.A.
 Herald, Christopher M., U.S.A.
 Herb, Rebecca A., U.S.A.
 Herfort, Wolfgang, Austria
 Herman, Eugene A., U.S.A.
 Herman, Richard H., U.S.A.
 Hermiller, Susan M., U.S.A.
 Hernandez, Eugenio, Spain
 Hernandez-Lamonedá, Luis,
 Mexico
 Herron, David A., U.S.A.
 Hersh, Reuben, U.S.A.
 Herz, Carl S., Canada
 Hettling, Karl F., U.S.A.
 Heumos, Michael J., U.S.A.
 Heuzé, Danièle, France
 Hibi, Takayuki, Japan
 Hickling, Fred, U.S.A.
 Hida, Haruzo, Japan
 Hidaka, Fumio, Japan
 Higgins, Philip J., England
 Higginson, William C., Canada
 Higuchi, Teiichi, Japan
 Hilario, Richard A., U.S.A.
 Hilbert, Stephen R., U.S.A.
 Hildebrand, Craig E., U.S.A.
 Hildenbrand, Werner, Federal
 Republic of Germany
 Hill, C. Denson, U.S.A.
 Hill, Shirley A., U.S.A.
 Hill, Theodore P., U.S.A.
 Hilton, Peter J., U.S.A.
 Himonas, Alex T., U.S.A.
 Hindman, Neil, U.S.A.
 Hinson, Edward K., U.S.A.
 Hirai, Atsuhiko, Japan
 Hirano, Yasuyuki, U.S.A.
 Hirsche, Herbert R., Federal
 Republic of Germany
 Hironaka, Eriko, U.S.A.
 Hironaka, Heisuke, U.S.A.
 Hirsch, Michael D., U.S.A.
 Hirsch, Morris W., U.S.A.
 Hirzebruch, Friedrich, Federal
 Republic of Germany
 Hitotsumatsu, Shin, Japan
 Ho, Chi-Fai, U.S.A.
 Ho, Chung-wu, U.S.A.
 Hoagland, Gordon W., U.S.A.
 Hobart, Sylvia A., U.S.A.
 Hochstadt, Harry, U.S.A.
 Hochster, Melvin, U.S.A.
 Hochwald, Scott H., U.S.A.
 Hodgson, Bernard R., Canada
 Hodgson, Craig D., U.S.A.
 Hodgkinson, Ian M., England
 Hofer, Edwin T., U.S.A.
 Hoffman, David A., U.S.A.
 Hoffman, Jerome W., U.S.A.
 Hoffman, Kenneth M., U.S.A.
 Hoffman, Michael J., U.S.A.
 Hoffman, Peter N., Canada
 Hogbe-Nlend, Henri, France
 Hohti, Aarno, Finland
 Holbrook, John A., Canada
 Holden, Herbert L., U.S.A.
 Holevo, Alexander S., U.S.S.R.
 Holland, Samuel S., U.S.A.
 Holland, Susan M., England
 Holland, W. Charles, U.S.A.
 Holley, Ann D., U.S.A.
 Holley, Frieda K., U.S.A.
 Holley, Richard A., U.S.A.
 Holm, Per, Norway
 Holmann, Harald R. A.,
 Switzerland
 Holmay, Kathleen, U.S.A.
 Holte, John M., U.S.A.
 Holtz, Marc, U.S.A.
 Hommel, Uwe G. M., Federal
 Republic of Germany
 Honda, Kinya, Japan
 Hoo, Cheons S., Canada
 Hood, J. Myron, U.S.A.
 Hood, Rodney T., U.S.A.
 Hooker, John W., U.S.A.
 Hörmander, Lars V., Sweden
 Hornung, Ulrich,
 Federal Republic of Germany
 Hou, Shui-Hung, Hong Kong
 Houh, Chong-Shi, U.S.A.
 Householder, James E., U.S.A.
 Houston, Johnny L., U.S.A.
 Houzel, Christian J., France
 Howard, Edgar J., U.S.A.
 Howard, Eric J., U.S.A.
 Howe, Vernon W., U.S.A.
 Hoyt, William L., U.S.A.
 Hrushovski, Ehud, U.S.A.
 Hsiang, Wu-chung, U.S.A.
 Hsiang, Wu-Teh, U.S.A.
 Hsiang, Wu-Yi, U.S.A.
 Hsieh, Po-Fang, U.S.A.
 Hsieh, Simon C., China-Taiwan
 Hsiung, Chuan-Chih, U.S.A.
 Hu, Guo-Ding, People's
 Republic of China
 Hu, Thakyin, China-Taiwan
 Huang, Douglas H. C., U.S.A.
 Huang, Ming-Deh, U.S.A.
 Huang, Qichang, People's
 Republic of China
 Hubbard, Elaine M., U.S.A.
 Hubbard, John H., U.S.A.
 Huber, Greg, U.S.A.
 Huber, Jay H., U.S.A.
 Hudson, Donna L., U.S.A.
 Huehn, Kempton L., U.S.A.
 Hughes, Daniel E., U.S.A.
 Hughes, John F., U.S.A.
 Hui, Stefan Y., U.S.A.
 Huibregtse, Mark E., U.S.A.
 Humke, Paul D., U.S.A.
 Humphreys, Edward J., England
 Humphreys, James E., U.S.A.
 Hunsaker, Worthen N., U.S.A.
 Hunt, Fern Y., U.S.A.
 Huotari, Robert H., U.S.A.
 Hurd, Albert E., Canada
 Hurder, Steven E., U.S.A.
 Hurtubise, Jacques, Canada
 Hurwitz, Carol M., U.S.A.
 Husch, Lawrence S., U.S.A.
 Hussain, Mansour G., Kuwait
 Huynh, Hsueh-ling, U.S.A.
 Hwang Jun Shung, China-
 Taiwan
 Hwang, Tea-Yuan, China-
 Taiwan
 Hyers, Donald H., U.S.A.
 Hynard, Edward O., U.S.A.
 Iarrobino, Anthony A., U.S.A.
 Ibisch, Horst, France
 Ibor-Latre, Luis A., Spain
 Ignjatović, Aleksandar Dj.,
 U.S.A.
 Ihara, Yasutaka, Japan
 Ikeda, Kazumasa, Japan
 Ikeshoji, Kiyoshi, Japan
 Illner, Reinhard, Canada
 Imai, Chuichi, Japan
 Im Hof, Hans C., Switzerland
 Ingram, Steven K., U.S.A.
 Ingrassia, Michael A., U.S.A.
 Ion, Patrick D. F., U.S.A.
 Iozzi, Alessandra, Italy
 Iraniparast, Nezam, U.S.A.
 Isac, George, Canada
 Ischebeck, Friedrich, Federal
 Republic of Germany
 Ishiguro, Kenshi, U.S.A.
 Ishihara, Shigeru, Japan
 Ishii, Shihoko, Japan
 Ishikawa, Fumie, Japan
 Isik, Nilgün, Turkey
 Ismail, Mourad E. H., U.S.A.
 Itô, Kiyosi, U.S.A.
 Ito, Ryuichi, Japan
 Ito, Yoshihiko, Japan
 Ito, Yoshitada, Japan

- Itoh, Mitsuhiro, U.S.A.
 Itzkowitz, Gerald L., U.S.A.
 Ivrii, Victor, U.S.S.R.
 Iwai, Toshihiro, Japan
 Iwaniec, Henryk, U.S.A.
 Izuchi, Keiji, Japan
 Izuchi, Yuko, Japan
 Jablow, Eric R., U.S.A.
 Jabon, David C., U.S.A.
 Jackson, Bradley W., U.S.A.
 Jacob, John P., U.S.A.
 Jacobowitz, Howard, U.S.A.
 Jaeger, Arno, Federal
 Republic of Germany
 Jafarian, Ali-Akbar, Iran
 Jaffe, Arthur M., U.S.A.
 Jaffe, David B., U.S.A.
 Jagy, William C., U.S.A.
 Jain, Atul, U.S.A.
 Jakobsche, Włodzimierz,
 Poland
 Jakobson, M. V., U.S.S.R.
 Jambor, Paul E., U.S.A.
 James, David M., U.S.A.
 James, Donald G., U.S.A.
 James, Ioan M., U.S.A.
 James, Robert C., U.S.A.
 Jamison, James E., U.S.A.
 Janke, Steven J., U.S.A.
 Janson, Barbara J., U.S.A.
 Janusz, Gerald J., U.S.A.
 Jardine, Rick, Canada
 Järvinen, Richard D., U.S.A.
 Jayne, J. E., England
 Jenab, Albert, U.S.A.
 Jenda, Overtoun M., Botswana
 Jenkins, Joe W., U.S.A.
 Jerison, David S., U.S.A.
 Jerison, Meyer, U.S.A.
 Jerrard, Richard P., U.S.A.
 Ji, Ronghui, U.S.A.
 Jiang, Boju, People's
 Republic of China
 Jiang Erxiong, People's
 Republic of China
 Jijitchenko, Alexei B., U.S.S.R.
 Jimenez, Pedro, Spain
 Jiménez Pozo, M. A., Cuba
 Johansson, Bo I., Sweden
 Johnson, Barry E., England
 Johnson, David, England
 Johnson, Gene D., U.S.A.
 Johnson, Gerald W., U.S.A.
 Johnson, Guy, U.S.A.
 Johnson, Harold H., U.S.A.
 Johnson, James T., U.S.A.
 Johnson, Keith P., Canada
 Johnson, Kenneth W., U.S.A.
 Johnson, R. Wells, U.S.A.
 Johnson, Roy A., U.S.A.
 Johnson, Steven D., U.S.A.
 Johnstone, Peter T., England
 Jolly, Bret J., U.S.A.
 Jones, Charles V., U.S.A.
 Jones, Douglas S., Scotland
 Jones, Earl P., U.S.A.
 Jones, Harold T., U.S.A.
 Jones, John D. S., England
 Jones, John, U.S.A.
 Jones, Peter W., U.S.A.
 Jones, Theodore W., U.S.A.
 Jones, Vaughan F. R., U.S.A.
 Jonker, Leo B., Canada
 Jonsson, Alf, Sweden
 Jonsson, Thordur, Iceland
 Jorgensen, Palle E. T., U.S.A.
 Jost, Juergen, Federal
 Republic of Germany
 Juister, Barbara J., U.S.A.
 Julg, Pierre, France
 Julianelle, Anthony, U.S.A.
 Jurca, Dan, U.S.A.
 Kachroo, Pandit Dilaram,
 Canada
 Kadison, Lars D., U.S.A.
 Kadison, Richard V., U.S.A.
 Kaftal, Victor G., U.S.A.
 Kågesten, Owe K., Sweden
 Kahan, W. M., U.S.A.
 Kahane, Jean-Pierre, France
 Kahn, Bruno, France
 Kahn, Peter J., U.S.A.
 Kahn, Steve, U.S.A.
 Kairies, Hans-Heinrich,
 Federal Republic of
 Germany
 Kaiser, Uwe, Federal
 Republic of Germany
 Kaji, Hajime, Japan
 Kajikawa, Yuji, Japan
 Kakehashi, Tetsujiro, Japan
 Kakutani, Shizuo, U.S.A.
 Kalhoff, Franz B., Federal
 Republic of Germany
 Kalin, Robert, U.S.A.
 Kalisch, Gerhard K., U.S.A.
 Kalman, John A., New Zealand
 Kalmbach, Gudrun, Federal
 Republic of Germany
 Kamber, Franz W., U.S.A.
 Kambule, Matthew T., U.S.A.
 Kaminker, Jerry A., U.S.A.
 Kaminsky, Kenneth S., U.S.A.
 Kammerer, Jean-Bertrand,
 France
 Kamowitz, Herbert, U.S.A.
 Kamran, Niky, Canada
 Kane, Jonathan M., U.S.A.
 Kaneko, Tetsuo, Japan
 Kanenobu, Taizo, Japan
 Kania-Bartoszynska, Joanna
 M., Poland
 Kanie, Yukihiro, Japan
 Kannappan, P. L., Canada
 Kanold, Hans-Joachim,
 Federal Republic of Germany
 Kantor, Jean-Michel, France
 Kaplan, Edward L., U.S.A.
 Kaplan, Russell M., U.S.A.
 Kaplan, Wilfred, U.S.A.
 Kaplansky, Irving, U.S.A.
 Kappagantula, Savithri, U.S.A.
 Karamzin, Youri N., U.S.S.R.
 Karbe, Manfred J.,
 Federal Republic of Germany
 Karel, Martin L., U.S.A.
 Karlsson, Thomas N., Sweden
 Karmarkar, Narendra, U.S.A.
 Karoubi, Max, France
 Karp, Leon, U.S.A.
 Karp, Richard M., U.S.A.
 Karpinski, Marek M., Federal
 Republic of Germany
 Karrer, Guido, Switzerland
 Karush, William, U.S.A.
 Kasdan, John M., U.S.A.
 Kass, Steve, Canada
 Kassab, John Y., Wales
 Kasue, Atsushi, Japan
 Kato, Goro, U.S.A.
 Kato, Mitsuyoshi, Japan
 Katō Sadao, Japan
 Kato, Tosio, U.S.A.
 Katsuura, Hidefumi, U.S.A.
 Katz, Arline S., U.S.A.
 Katz, Irving J., U.S.A.
 Katz, Richard, U.S.A.
 Katz, Victor J., U.S.A.
 Katzoff, Myron J., U.S.A.
 Kaufman, Allan N., U.S.A.
 Kaufman, Arthur, U.S.A.
 Kaufmann-Buehler, Walter,
 U.S.A.
 Kaul, Saroop K., Canada
 Kavianpour, Alireza, Iran
 Kawaguchi, Michiaki, Japan
 Kawahigashi, Yasuyuki, U.S.A.
 Kawashima, Toshio, Japan
 Kazhdan, David, U.S.A.
 Kearnes, Keith A., U.S.A.
 Kearton, Cherry, England
 Keats, Reynold G., Australia
 Kechris, Alexander S., U.S.A.
 Keen, Linda, U.S.A.
 Keene, Frederick W., U.S.A.
 Kegel, Otto H.,
 Federal Republic of Germany
 Keller, Heinrich H., Switzerland
 Kelley, John L., U.S.A.
 Kellogg, Franklin R., U.S.A.

- Kellogg, Mary, U.S.A.
 Kelly, Alice J., U.S.A.
 Kelly, John B., U.S.A.
 Kelly, Leroy M., U.S.A.
 Kelly, Michael R., U.S.A.
 Kemnitz, Arnfried, Federal
 Republic of Germany
 Kemp, Paula Ann, U.S.A.
 Kempf, George R., U.S.A.
 Kenderov, Petar, Bulgaria
 Kenig, Carlos E., U.S.A.
 Kenschaft, Patricia, U.S.A.
 Keown, Ray, U.S.A.
 Kérchy, László, Hungary
 Kerckhoff, Steven P., U.S.A.
 Kersten, Ina, Federal
 Republic of Germany
 Kerzman, Norberto L., U.S.A.
 Kestenband, B., U.S.A.
 Keum, Jonghae, U.S.A.
 Key, Jennifer D., England
 Keynes, Harvey B., U.S.A.
 Khalifah, Mohammed F., Oman
 Khalimsky, Efim D., U.S.A.
 Khaliq, A. Q. M., Bahrain
 Kharaghani, Hadi, Canada
 Khare, Shabd S., India
 Khatri, Dharam C., U.S.A.
 Khelladi, Abdelkader, Algeria
 Khosrovshahi, Gholamreza B.,
 Iran
 Kidwell, Mark E., U.S.A.
 Kidwell, Peggy A., U.S.A.
 Kim, Dohan, South Korea
 Kim, Dong-Gyom, U.S.A.
 Kim, Myong-Hi, U.S.A.
 Kim, Yonggu, U.S.A.
 Kim, Yonne Mi, U.S.A.
 Kimura, Naoki, U.S.A.
 King, Henry C., U.S.A.
 King, Jonathan L., U.S.A.
 King, Oliver H., England
 Kipps, Thomas C., U.S.A.
 Kirby, Robion C., U.S.A.
 Kirk, Ronald B., U.S.A.
 Kirmayer, Greg, U.S.A.
 Kishore, Masao, U.S.A.
 Kister, James M., U.S.A.
 Kister, Jane E., U.S.A.
 Klaasen, Gene A., U.S.A.
 Klass, Michael J., U.S.A.
 Klawe, Maria, U.S.A.
 Kleiner, Bruce A., U.S.A.
 Knapp, Anthony W., U.S.A.
 Knill, Ronald J., U.S.A.
 Knoebel, Arthur, U.S.A.
 Knopp, Marvin I., U.S.A.
 Kobayashi, Keiko, Japan
 Kobayashi, Masanori, Japan
 Kobayashi, Shoshichi, U.S.A.
 Kobayashi, Tsuyoshi, Japan
 Kobayashi, Yuji, Japan
 Koçak, Cevdet, Turkey
 Kocak, Huseyin, U.S.A.
 Koch, Helmut V., German
 Democratic Republic
 Koch, Lisa K., U.S.A.
 Kochendorfer, Anne, U.S.A.
 Kochman, Stanley O., Canada
 Koenig, Gerhard, Federal
 Republic of Germany
 Kohlmayr, Gerhard F., U.S.A.
 Kohn, Joseph J., U.S.A.
 Kohno, Toshitake, Japan
 Koike, Kazuhiko, Japan
 Koiller, Jair, U.S.A.
 Kolk, Johan A., Netherlands
 Kolodner, Ignace I., U.S.A.
 Komatsu, Hikosaburo, Japan
 Komatsu, Yûsaku, Japan
 Komjath, Peter, Hungary
 Kommel, Helene J., Canada
 Kon, Mark A., U.S.A.
 Konate, Lamine, Ivory Coast
 Konhauser, Joseph D. E.,
 U.S.A.
 Koranyi, Adam, U.S.A.
 Korenblum, Boris, U.S.A.
 Korevaar, Jacob, Netherlands
 Korzan, Dale M., U.S.V.I.
 Kosachevskya, Yelena,
 U.S.S.R.
 Kosaki, Hideki, Japan
 Koshi, Shozo, Japan
 Kosinski, Antoni A., U.S.A.
 Koslowski, Jürgen, U.S.A.
 Kosmala, Witold A. J., U.S.A.
 Kostlan, Eric J., U.S.A.
 Kota, Osamu, Japan
 Koua, Konin, Ivory Coast
 Kovacec, Alexander, Austria
 Kovács, László G., Australia
 Kövari, Thomas, England
 Kovářík, Zdislav V., Canada
 Kövesi, Endre H., Austria
 Kozlov, V. V., U.S.S.R.
 Krabbe, Gregers L., U.S.A.
 Krairojananan, Sompop,
 Thailand
 Kramer, Jurg, U.S.A.
 Krantz, Steven G., U.S.A.
 Kranz, Przemot T., U.S.A.
 Kratz, Werner, Federal
 Republic of Germany
 Kreider, Donald L., U.S.A.
 Kreith, Kurt, U.S.A.
 Krener, Arthur J., U.S.A.
 Kreuzman, Mary J., U.S.A.
 Kreyszig, Erwin, Canada
 Krichevsky, Rafail Evseevič,
 U.S.S.R.
 Krieger, Henry A., U.S.A.
 Krieger, Wolfgang J., Federal
 Republic of Germany
 Kristensen, Leif, Denmark
 Krom, Mel, U.S.A.
 Krop, Leonard, U.S.A.
 Kruzhilin, N. G., U.S.S.R.
 Kryazhinskii, Arkadii V.,
 U.S.S.R.
 Krylov, N. V., U.S.S.R.
 Ku, Hsu-Tung, U.S.A.
 Ku, Mei-Chin H., U.S.A.
 Kubelka, Richard P., U.S.A.
 Kubilius, Jonas P., U.S.S.R.
 Kufner, Alois, Czechoslovakia
 Kuhn, Nicholas J., U.S.A.
 Kuiper, Nicolaas H., France
 Kuku, Aderemi O., Nigeria
 Kumo, Koichiro, Japan
 Kunita, Hiroshi, Japan
 Kunoff, Sharon M., U.S.A.
 Kunze, Ray A., U.S.A.
 Kuo, Tsang-Hai, China-Taiwan
 Kuperberg, Gregory J.,
 U.S.A.
 Kuperberg, Włodzimierz, U.S.A.
 Kupka, Ivan A. K., France
 Kurke, Herbert, German
 Democratic Republic
 Kurshan, Robert P., U.S.A.
 Kurss, Herbert, U.S.A.
 Kurtzke, John F., U.S.A.
 Kutzko, Philip C., U.S.A.
 Kuwabara, Ruishi, Japan
 Kwak, Nosup, U.S.A.
 Kyle, Joseph, England
 Kyner, Walter T., U.S.A.
 Labarca, Rafael, Chile
 Laca, Marcelo E., U.S.A.
 Lacampagne, Carole B., U.S.A.
 Lacey, H. Elton, U.S.A.
 Lachlan, Alistair H., Canada
 Lacomba, Ernesto A., Mexico
 Laczkovich, Miklos, U.S.A.
 LaDuke, Jeanne, U.S.A.
 Ladwig, Jack, U.S.A.
 Lafferty, John D., U.S.A.
 Laffey, Thomas J., Ireland
 Lagarias, Jeffrey C., U.S.A.
 Lai, Hang-Chin, China-Taiwan
 Lai, King F., Hong Kong
 Lai, Michael Y., U.S.A.
 Lalli, Bikkar Singh, Canada
 Lam, Tsit-Yuen, U.S.A.
 Lamb, John F., U.S.A.
 Lambert, William M., Costa
 Rica
 Lamken, Esther R., U.S.A.

- Lamoureux, Michael P., U.S.A.
 Lancaster, Kirk E., U.S.A.
 Lance, E. Christopher, England
 Lander, Felix I., U.S.A.
 Lane, Mark T., U.S.A.
 Lane, Mary C., U.S.A.
 Lanford, Oscar E., France
 Lang, William C., U.S.A.
 Lang, William E., U.S.A.
 Lange, John E., U.S.A.
 Lange, Lester H., U.S.A.
 Langevin, Michel, France
 Lansdon, Robert S., U.S.A.
 Lapidus, Arnold, U.S.A.
 Lapidus, Michel L., U.S.A.
 Lappan, Peter A., U.S.A.
 Lardy, Lawrence J., U.S.A.
 Laroco, Leonardo A., U.S.A.
 Larsen, Michael J., U.S.A.
 Larson, Jean A., U.S.A.
 Larson, Loren C., U.S.A.
 Larue, David M., U.S.A.
 Laskowski, Michael C., U.S.A.
 Lassner, Gerd, German
 Democratic Republic
 Latham, Geoff A., U.S.A.
 Lau, Anthony To-Ming,
 Canada
 Lau, Kee-Wai, Hong Kong
 Laubenbacher, Reinhard C.,
 U.S.A.
 Laumon, Gerard H., France
 Laursen, Kjeld B., Denmark
 Laville, Guy A., France
 Lavrov, Igor, U.S.S.R.
 Law, Peter R., U.S.A.
 Lawrence, John W., Canada
 Lawson, James W., U.S.A.
 Lawson, Terry C., U.S.A.
 Lax, Peter D., U.S.A.
 Laywine, Charles F., Canada
 Lazar, Aldo J., Israel
 Lazarov, Connor, U.S.A.
 Lazarus, Andrew J., U.S.A.
 Lazarus, Susan, U.S.A.
 Lazebnik, Felix, U.S.A.
 Lazerson, Earl E., U.S.A.
 Lbekkouri, Aboubakr, U.S.A.
 Lê, Dung Trang, France
 Leader, Solomon, U.S.A.
 Leary, Christopher C., U.S.A.
 Leary, Kevin, U.S.A.
 LeBlanc, Emile A., U.S.A.
 Leborgne, Daniel, France
 Lebow, Arnold, U.S.A.
 LeBrun, Claude R., U.S.A.
 Lecomte, Florence C., France
 Lee, John M., U.S.A.
 Lee, Jong P., U.S.A.
 Lee, King Tak, Malaysia
 Lee Peng-Yee, Singapore
 Lee, Rebecca A., U.S.A.
 Lee, Richard W. M., Canada
 Lee, Sin-Min, U.S.A.
 Lee, Tzong-Yow, China-Taiwan
 Lee, Will Y., U.S.A.
 Leech, Jonathan E., U.S.A.
 Leep, David B., U.S.A.
 Lefton, Phyllis, U.S.A.
 Leggett, Anne, U.S.A.
 Legrand, Stephen, U.S.A.
 Lehman, R. Sherman, U.S.A.
 Lehrer, Gus I., Australia
 Lehrer-Ilamed, Yehiel, Israel
 Lehto, Olli E., Finland
 Leibler, Richard A., U.S.A.
 Leichtweiss, Kurt, Federal
 Republic of Germany
 Leifman, Lev J., U.S.A.
 Leighly, Joseph, U.S.A.
 Lempert, László, Hungary
 Lenhart, Suzanne M., U.S.A.
 Lennox, John C., Wales
 Lenstra Jr., H. W., Netherlands
 Lepowsky, James, U.S.A.
 Lerman, Manuel, U.S.A.
 Leruste, Christian P., France
 Lesieur, Leonce E., France
 Leskinen, Anne L., U.S.A.
 Leslie, Joshua A., U.S.A.
 Leung, Ka Hin, U.S.A.
 Leung, Kar-Koi, China-Taiwan
 Leung, Nal Chung, Hong Kong
 Leung, Pui-Fai Fred, Singapore
 Levenberg, Norman, U.S.A.
 LeVeque, William J., U.S.A.
 Levesque, Claude, Canada
 Levi, Mark, U.S.A.
 Levine, Marc N., U.S.A.
 Lewis, Albert C., Canada
 Lewis, Donald J., U.S.A.
 Lewis, George M., U.S.A.
 Lewis, L. Gaunce, U.S.A.
 Lewis, Richard P., England
 Lewis, Wayne, U.S.A.
 Lewy, Hans, U.S.A.
 Li Chi-Kwong, Hong Kong
 Li, Jing, U.S.A.
 Li Kam-Yuk, Hong Kong
 Li, Ke-Zheng, U.S.A.
 Li, Kin Y., U.S.A.
 Li, Ming-Sun, U.S.A.
 Li, Wen-Ching, U.S.A.
 Li, Zezhi, People's Republic
 of China
 Liao, Shao Tao, People's
 Republic of China
 Libgober, Anatoly S., U.S.A.
 Libis, Carl A., U.S.A.
 Lick, Don R., U.S.A.
 Lickorish, W. B. Raymond,
 England
 Lidl, Rudolf, Australia
 Lieberman, David I., U.S.A.
 Lieberman, Gary M., U.S.A.
 Lieberman, Henry S., U.S.A.
 Liggett, Thomas M., U.S.A.
 Ligh, Steve, U.S.A.
 Lim, Chong Hai, U.S.A.
 Lim, Chong-keang, Malaysia
 Lin, Charles S. C., U.S.A.
 Lin, Charlotte, U.S.A.
 Lin, En-Bing, U.S.A.
 Lin, Florence J., U.S.A.
 Lin, James P., U.S.A.
 Lin, Kai-Ching, U.S.A.
 Lin, Shine-Min, U.S.A.
 Lin, Wen-Hsiung, China-
 Taiwan
 Lin, Xiao-Song, U.S.A.
 Lin, Zhen-sheng, People's
 Republic of China
 Lind, Douglas A., U.S.A.
 Lindberg, Karl J., U.S.A.
 Linfoot, Joyce J., England
 Ling, San, U.S.A.
 Linton, Fred E. J., U.S.A.
 Lipkin, Leonard J., U.S.A.
 Lipschutz, Seymour, U.S.A.
 Lipshutz, Robert J., U.S.A.
 Lipsich, H. David, U.S.A.
 Lipsman, Ronald L., U.S.A.
 Little, John B., U.S.A.
 Little, Julia A., U.S.A.
 Liu Fon-Che, China-Taiwan
 Liu, Shih-Chao, China-Taiwan
 Liu Shize, People's Republic
 of China
 Liu, Yingming, People's
 Republic of China
 Liu, Zhong-Dong, U.S.A.
 Liulevicius, Arunas L., U.S.A.
 Llosa, Betsy, U.S.A.
 Llovet, Juan, Spain
 Lluís, Emilio, Mexico
 Loase, John F., U.S.A.
 Lobel, Perry D., U.S.A.
 Loewy, Raphael, Israel
 Lofquist, George W., U.S.A.
 Lojasiewicz, W., U.S.A.
 London, David, Israel
 Long, Darren, U.S.A.
 Longyear, Judith Q., U.S.A.
 Lopes, Artur O., Brazil
 Lopes, Thomas R., U.S.A.
 López de Medrano, Santiago,
 Mexico
 Lorch, Lee A., Canada
 Lorentz, Rudolph A., Federal
 Republic of Germany
 Lorenzini, Dino, U.S.A.

- Loring, Terry A., U.S.A.
 Lotto, Benjamin A., U.S.A.
 Lounesto, Pertti, Finland
 Lovasz, Laszlo, Hungary
 Loveys, James G., Canada
 Lowengrub, John, U.S.A.
 Lozano, Maria T., Spain
 Lu, Jiang-Hua, U.S.A.
 Lubkin, Saul, U.S.A.
 Lucas, William F., U.S.A.
 Luecke, John E., U.S.A.
 Luk, Hing-Sun, Hong Kong
 Luke, Stanley D., U.S.A.
 Lumer, Gunter, Belgium
 Lumon, Françoise M., France
 Luna, Dominique, France
 Luna, George W., U.S.A.
 Lundell, Albert T., U.S.A.
 Lutterodt, Clement H., U.S.A.
 Luukkainen, Jouni, Finland
 Luxemburg, Wilhelmus A., U.S.A.
 Lyons, Russell D., U.S.A.
 Ma, Daowei, U.S.A.
 Ma, Lawrence K., U.S.A.
 Ma, Siu Lun, Hong Kong
 Ma, Xiaoyun, U.S.A.
 Mabuchi, Toshiaki, Federal Republic of Germany
 Mac Donnell, John J., U.S.A.
 Macdonald, Ian D., Scotland
 MacDonald, James H., U.S.A.
 MacDonnell, Joseph F., U.S.A.
 Macey, Wade T., U.S.A.
 Machedon, Matei, U.S.A.
 Macintyre, Angus J., England
 Mackenzie, Kirill, Australia
 Mackenzie, Kirsten, England
 MacKenzie, Robert E., U.S.A.
 Madan, Manohar L., U.S.A.
 Madan, Ved P., Canada
 Maddux, Roger D., U.S.A.
 Maehara, Kazuhisa, Japan
 Magarian, Elizabeth A., U.S.A.
 Magenes, Enrico, Italy
 Maggioni, Renato, Italy
 Magid, Andy R., U.S.A.
 Magid, Martin A., U.S.A.
 Magidor, Menachem, Israel
 Magnus, Arne, U.S.A.
 Magnus, John B., U.S.A.
 Magnus, Peggy H., U.S.A.
 Magnuson, Alan W., U.S.A.
 Magnússon, Jón I., Iceland
 Mahaney, Lou Ann, U.S.A.
 Mahdavi-Hezavehi, M., Iran
 Mahmoodian, Ebadollah S., Iran
 Mahony, Louis M., U.S.A.
 Maier, Franz, U.S.A.
 Maier, Helmut, U.S.A.
 Majumdar, Samir R., Canada
 Makar-Limanov, Leonid, U.S.A.
 Makar-Limanov, Olga, U.S.A.
 Makarov, N. G., U.S.S.R.
 Mäkeläinen, Tuulikki, Finland
 Makky, Sadia M., Iraq
 Malcolmson, Peter, U.S.A.
 Malecki, Sophie, France
 Malek, Massoud, U.S.A.
 Maleki, Amir A., U.S.A.
 Malliavin, Paul, France
 Malm, Donald E. G., U.S.A.
 Malouf, Janice, U.S.A.
 Maltsiniotis, Georges, France
 Malyshev, Igor, U.S.A.
 Mammitzsch, Volker, Federal Republic of Germany
 Mamourian, Ahmad, Iran
 Manders, Kenneth L., U.S.A.
 Manderscheid, David C., U.S.A.
 Manferdelli, John L., U.S.A.
 Manfredi, Juan J., U.S.A.
 Mangano, Gian Carlo, U.S.A.
 Manin, Avwohm, Israel
 Manin, Yuri, U.S.S.R.
 Mann, Benjamin M., U.S.A.
 Mann, Larry N., U.S.A.
 Manocha, H. L., India
 Manolov, Spas M., Bulgaria
 Manougian, Edward, U.S.A.
 Manushin, Victor, U.S.S.R.
 Marathe, Kishore B., U.S.A.
 Marchenko, Vladimir, U.S.S.R.
 Marcos, Eduardo N., U.S.A.
 Marcotte, Odile, Canada
 Marcum, Howard J., U.S.A.
 Marcus, Diana, U.S.A.
 Marcus, Matthew A., U.S.A.
 Margalef-Roig, Juan, Spain
 Margarit, Alejandro F., Spain
 Margolis, Stuart W., U.S.A.
 Margulies, Caryl A., U.S.A.
 Margulies, William G., U.S.A.
 Marini, Luisa D., Italy
 Marino, Mario, Italy
 Maritz, Pieter, South Africa
 Marker, David E., U.S.A.
 Marsh, Marcus M., U.S.A.
 Marshall, Donald E., U.S.A.
 Martellotti, Anna, Italy
 Martin, Donald A., U.S.A.
 Martin, E. Dale, U.S.A.
 Martin, Gary A., U.S.A.
 Martin, Nathaniel F. G., U.S.A.
 Martineau, Robert P., England
 Martio, Olli T., Finland
 Maruo, Osamu, Japan
 Maruyama, Fumitsuna, Japan
 Marxen, Donald J., U.S.A.
 Masahara, Kazuhisa, Japan
 Massé, Jean-Claude, Canada
 Massell, Paul B., U.S.A.
 Massera, Jose L., Uruguay
 Masuda, Tamezo, Japan
 Masuda, Tetsuya, France
 Mate, Attila, U.S.A.
 Mather, John N., U.S.A.
 Mathieu, Martin, Federal Republic of Germany
 Mathieu, Philippe, Belgium
 Matic, Gordana, U.S.A.
 Matsue, Hirofumi, Japan
 Matsuki, Kenji, Japan
 Matsumoto, Koji, Japan
 Matsumoto, Shigenori, Japan
 Matsumura, Hideyuki, Japan
 Matsuoaka, Cho-Ichiro, Japan
 Matsushita, Yasuo, Japan
 Matsuzawa, Junichi, Japan
 Matthes, Klaus, German Democratic Republic
 Mattison, Robert, U.S.A.
 Maude, Ronald, England
 Maumary, Serge, Switzerland
 Mauro, Fabrizio, Italy
 Mawhin, Jean, Belgium
 Mawyer, Farley, U.S.A.
 Maxfield, John E., U.S.A.
 Maxoudov, Faramaz, U.S.S.R.
 Maxwell, Christine, U.S.A.
 Maxwell, James W., U.S.A.
 May, Eleanor, U.S.A.
 May, J. Peter, U.S.A.
 May, Michael K., U.S.A.
 Mayer, David, Mexico
 Mayer-Wolf, Eddy, Israel
 Mazur, Barry C., U.S.A.
 Mazzeo, Rafe R., U.S.A.
 McArthur, Charles W., U.S.A.
 McAuley, Van A., U.S.A.
 McBrien, Vincent O., U.S.A.
 McCallum, William G., U.S.A.
 McCarthy, John D., U.S.A.
 McCarthy, John E., U.S.A.
 McCartin, Brian J., U.S.A.
 McCaughan, Dennis J., New Zealand
 McCleary, John H., U.S.A.
 McCray, Patrick D., U.S.A.
 McCrory, Clint G., U.S.A.
 McCrudden, Michael, England
 McCullough, Darryl J., U.S.A.
 McCurley, Kevin S., U.S.A.
 McDaniel, Andrew, U.S.A.
 McDonald, Bernard R., U.S.A.
 McDonald, Gary C., U.S.A.

- McDonald, Gerard, U.S.A.
 McDonald, John N., U.S.A.
 McEvoy, Brian F., U.S.A.
 McFaden, Harold, U.S.A.
 McGehee, Richard P., U.S.A.
 McKay, James H., U.S.A.
 McKay, John, Canada
 McKean, Henry P., U.S.A.
 McKenzie, Ralph N., U.S.A.
 McLaren, Christine E., U.S.A.
 McLaughlin, Renate, U.S.A.
 McLean, Robert C., U.S.A.
 McMinn, Trevor J., U.S.A.
 McMullen, Curtis T., U.S.A.
 McNab, Robert B., U.S.A.
 McOwen, Robert C., U.S.A.
 McRae, George, U.S.A.
 McSwiggen, Patrick D., U.S.A.
 Meadows, Douglas S., U.S.A.
 Meakin, John C., U.S.A.
 Meeks, William H., U.S.A.
 Megginson, Robert E., U.S.A.
 Megibben, Charles K., U.S.A.
 Mehri, Bahman, Iran
 Meier, David, Switzerland
 Meisters, Gary H., U.S.A.
 Mejia, Raymond, U.S.A.
 Mejtsky, George J., U.S.A.
 Mejzler, David, Israel
 Mekler, Alan H., Canada
 Méla, Jean-François, France
 Mellender, James W., U.S.A.
 Melo, Severino T. R., Brazil
 Melvin, Paul M., U.S.A.
 Mendelson, Elliott, U.S.A.
 Meredith, David B., U.S.A.
 Merkle, Milan J., Yugoslavia
 Merkurjev, A., U.S.S.R.
 Merris, Russell, U.S.A.
 Mertens, Jean-François,
 Belgium
 Merzbach, Uta C., U.S.A.
 Mesirov, Jill P., U.S.A.
 Mess, Geoffrey, Canada
 Messing, William, U.S.A.
 Meyer, Burnett, U.S.A.
 Meyer, Kenneth R., U.S.A.
 Meyer, Wolfgang T., Federal
 Republic of Germany
 Meyerhoff, Robert, U.S.A.
 Meyerowitz, Aaron D., U.S.A.
 Meyers, Leroy F., U.S.A.
 Micallef, Mario J., U.S.A.
 Michael, Mark, U.S.A.
 Michaelis, Walter J., U.S.A.
 Michnowicz, Theresa C., U.S.A.
 Mickelson, Jouko A., Finland
 Mihaljinec, Mirko, Yugoslavia
 Mihalik, Michael L., U.S.A.
 Mikami, Kentaro, U.S.A.
 Miklos, Dezso, Hungary
 Mikolás, Miklós, Hungary
 Mikulska, Margaret, U.S.A.
 Mikulski, Piotr W., U.S.A.
 Milcetic, John G., U.S.A.
 Miles, E. P., U.S.A.
 Milgram, Richard J., U.S.A.
 Miller, D. D., U.S.A.
 Miller, Haynes R., France
 Miller, John G., U.S.A.
 Miller, Keith, U.S.A.
 Miller, Sanford S., U.S.A.
 Miller, Victor S., U.S.A.
 Millett, Kenneth C., U.S.A.
 Millman, Richard S., U.S.A.
 Milman, Vitali, Israel
 Milne, Stephen C., U.S.A.
 Milnes, Paul, Canada
 Milnor, John W., U.S.A.
 Miloslavsky, George
 Constantinovich, U.S.A.
 Mináč, Ján, Canada
 Minami, Norihiko, U.S.A.
 Minc, Henryk, U.S.A.
 Mingo, James A., Canada
 Min-oo, Maung, Canada
 Miranda, Isabel, Spain
 Miranker, Willard L., U.S.A.
 Mirsky, Norman D., U.S.A.
 Misaki, Norihiro, Japan
 Mishra, Sudhakara, U.S.A.
 Misra, Arvind K., Nigeria
 Mitchell, Josephine M., U.S.A.
 Mitchem, John A., U.S.A.
 Mitropolski, Iouri, U.S.S.R.
 Miwa, Tetsuji, Japan
 Miyajima, Shizuo, Japan
 Miyazaki, Morishige, Japan
 Mizohata, Sigeru, Japan
 Mizutani, Tadayoshi, Japan
 Mockenhaupt, Gerd, Federal
 Republic of Germany
 Moishezon, Boris, U.S.A.
 Moll, Victor H., U.S.A.
 Mollin, Richard A., Canada
 Montalto, Barbara J., U.S.A.
 Montanari, Daniele, U.S.A.
 Montaudouin, Yves, U.S.A.
 Montejano, Luis, Mexico
 Montgomery, Joe M., U.S.A.
 Montgomery, M. Susan, U.S.A.
 Montgomery, Peter L., U.S.A.
 Moody, John A., U.S.A.
 Moore, Calvin C., U.S.A.
 Moore, Charles N., U.S.A.
 Moore, John D., U.S.A.
 Moore, Richard A., U.S.A.
 Moore, Robert L., U.S.A.
 Morais, Ricardo, Brazil
 Morales, Leonel, Guatemala
 Morales, Pedro A., Canada
 Moran, Gadi W., Israel
 Morawetz, Cathleen S., U.S.A.
 Moreira, Manuel R. F.,
 Portugal
 Morgan, Frank, U.S.A.
 Morgan, John W., U.S.A.
 Morgan, Kathryn A., U.S.A.
 Morgenstern, Jacques, France
 Morimoto, Masaharu, Japan
 Morisugi, Kaoru, Japan
 Moritz, Roger H., U.S.A.
 Moriwaki, Atsushi, Japan
 Moriyoishi, Hitoshi, Japan
 Morris, S. Brent, U.S.A.
 Morrison, David R., U.S.A.
 Morrison, Kent E., U.S.A.
 Morton, Hugh R., England
 Morton, Patrick, U.S.A.
 Moser, Jürgen, Switzerland
 Moser, Louise E., U.S.A.
 Moses, Michael F., U.S.A.
 Mosher, Lee D., U.S.A.
 Mosier, Ronald G., U.S.A.
 Moss, Larry S., U.S.A.
 Mostow, G. Daniel, U.S.A.
 Motupalli, Satyanarayana,
 U.S.A.
 Moura Neto, Francisco D.,
 U.S.A.
 Moy, Allen, U.S.A.
 Mrowka, Tomasz S., U.S.A.
 Muder, Douglas J., U.S.A.
 Mueller, James R., U.S.A.
 Mueller, Joseph E., U.S.A.
 Mugler, Dale H., U.S.A.
 Mühlbach, Günter W.,
 Federal Republic of Germany
 Mukerjee, Himadri K., India
 Mulase, Motohico, U.S.A.
 Mulcahy, Marjorie, U.S.A.
 Muldoon, Martin E., Canada
 Müller, Thomas H.,
 Switzerland
 Mulvey, Christopher J.,
 England
 Munkholm, Hans J., Denmark
 Murakami, Hitoshi, Japan
 Murasugi, Kunio, Canada
 Murphy, Paul F., U.S.A.
 Murray, Charles L., U.S.A.
 Murthy, M. P., India
 Myerson, Gerald, U.S.A.
 Myung, Hyo C., U.S.A.
 Naeini, Mohammad Kazem,
 Iran
 Naftalevich, Aaron, U.S.A.
 Nagahara, Takashi, Japan
 Nagaoka, Shoyu, Japan
 Nagasaka, Kenji, Japan

- Nagase, Masayoshi, Japan
 Nagata, Chigusa, Japan
 Nagata, Masayoshi, Japan
 Nagayama, Haruya, Japan
 Nainpally, Som A., Canada
 Nakai, Isao, England
 Nakamura, Yoshimasa, Japan
 Nakanishi, Shizu, Japan
 Nakashima, Katsuya, Japan
 Nakayama, Noboru, Japan
 Namba, Kanji, Japan
 Nance, Douglas W., U.S.A.
 Nara, Chiê, U.S.A.
 Narasimhan, Mudumbai S., India
 Narayanaswami, P. P., Canada
 Naroditsky, Vladimir, U.S.A.
 Nash, David H., U.S.A.
 Nashed, M. Z., U.S.A.
 Natarajan, Loki, U.S.A.
 Nathanson, Melvyn B., U.S.A.
 Nathanson, Weston I., U.S.A.
 Natsume, Toshikazu, Japan
 Natvig, Jon, Norway
 Nebres, Bienvenido F., Philippines
 Neff, John D., U.S.A.
 Neff, Mary M., U.S.A.
 Nejad Dehghan, Yadollah, Iran
 Nel, Louis D., Canada
 Nelson, David G., U.S.A.
 Nelson, George C., U.S.A.
 Ness, Harald M., U.S.A.
 Ness, Linda A., U.S.A.
 Neuberger, Barbara O., U.S.A.
 Neuberger, John W., U.S.A.
 Neumann, Bernhard H., Australia
 Neumann, Michael M., Federal Republic of Germany
 Neustadter, Siegfried F., U.S.A.
 Neuvonen, Timo, Finland
 Nevanlinna, Olavi, Finland
 Newberger, Stuart M., U.S.A.
 Newman, Michael F., Australia
 Newton, Paul K., U.S.A.
 Nezit, Pierre, Ivory Coast
 Ng, C. T., Canada
 Ng, Eugene K. S., U.S.A.
 Ng, Ho Kuen, U.S.A.
 Ng, K. C., U.S.A.
 Ngo, Viet, U.S.A.
 Nguyen Quang Do, Thong, France
 Nicas, Andrew J., Canada
 Nicholson, W. Keith, Canada
 Nico, William R., U.S.A.
 Nicolas, Anne-Marie, France
 Nicolas, Jean-Louis, France
 Nicolau, Monica, U.S.A.
 Nicolosi, Francesco, Italy
 Niehammer, Wilhelm, Federal Republic of Germany
 Niino, Kiyoshi, Japan
 Nijenhuis, Albert, U.S.A.
 Nikoltjeva-Hedberg, Margarita G., Sweden
 Nikulin, V. V., U.S.S.R.
 Nipp, Gordon L., U.S.A.
 Nishimori, Toshiyuki, Japan
 Nishimoto, Katsuyuki, Japan
 Nitsure, Nitin, India
 Niven, Ivan, U.S.A.
 Noble, Steven S., U.S.A.
 Nolder, Craig A., U.S.A.
 Nollet, Scott R., U.S.A.
 Norita, Sadataka, Japan
 Norman, Edward, U.S.A.
 Norman, R. Daniel, Canada
 Norman, Sven, Sweden
 Noronha, Maria Helena, Brazil
 Norrie, Katherine J., England
 Norris, John H., U.S.A.
 Norton, Alec, U.S.A.
 Norton, Douglas E., U.S.A.
 Nottrot, Roel, Netherlands
 Nower, Leon, U.S.A.
 Nummela, Eric C., England
 Ó Mathúna, Diarmuid, Ireland
 Obaid, Samih A., U.S.A.
 Oberlin, Daniel M., U.S.A.
 O'Brian, Nigel R., O' Australia
 Ó'Cairbre, Fiacre A., U.S.A.
 Ocneanu, Adrian, U.S.A.
 O'Connor, Vincent F., U.S.A.
 Odlyzko, Andrew M., U.S.A.
 Offner, Carl D., U.S.A.
 Ofman, Salomon, France
 Ogawa, Toshiyuki, Japan
 Ogg, Andrew P., U.S.A.
 Ogiso, Keiji, Japan
 Ogle, Crichton L., U.S.A.
 Ogus, Arthur E., U.S.A.
 Oh, Kyungho, U.S.A.
 Oh, Yong-Geun, U.S.A.
 Ohm, Jack E., U.S.A.
 Ohmiya, Mayumi, Japan
 Ohtsuka, Kohji, Japan
 Ojakian, Ronald S., U.S.A.
 Ojanguren, Manuel, Switzerland
 Okada, Masami, Japan
 Okazawa, Noboru, Japan
 Oldham, Janis M., U.S.A.
 Olech, Czesław, Poland
 Olevski, Alexandre M., U.S.S.R.
 Oliker, Vladimir, U.S.A.
 Olin, Robert F., U.S.A.
 Olsen, Catherine L., U.S.A.
 Olshen, Richard A., U.S.A.
 Olson, Dwight M., U.S.A.
 Olson, Loren D., Norway
 Olwell, Kevin D., U.S.A.
 O'Meara, O. Timothy, U.S.A.
 O'Neill, John D., U.S.A.
 Onneweer, Kees, U.S.A.
 Oppenheim, Joseph H., U.S.A.
 Orszag, Steven A., U.S.A.
 Ortiz, Carmen A., U.S.A.
 Osborn, J. Marshall, U.S.A.
 Osgood, Brad G., U.S.A.
 Osgood, Charles F., U.S.A.
 O'Shea, Donal B., U.S.A.
 Osofsky, Barbara L., U.S.A.
 Osserman, Robert, U.S.A.
 Osterburg, James, U.S.A.
 Otsuka, Kayo, Japan
 Otsuki, Nobukazu, Japan
 Otsuki, Tominosuke, Japan
 Ouedraogo, Georgette, Ivory Coast
 Ouzong, Samir, U.S.A.
 Ovchinnikov, Sergey V., U.S.A.
 Owa, Shigeyoshi, Japan
 Oyabu, Takashi, Japan
 Oyuke, Benjamin C., Kenya
 Ozeki, Michio, Japan
 Paalman-deMiranda, Aida B., Netherlands
 Pacheco, Nelson S., U.S.A.
 Pacifico, Maria J., Brazil
 Packel, Edward W., U.S.A.
 Pagani, Carlo D., Italy
 Page, Stanley S., Canada
 Page, Warren, U.S.A.
 Pai, Chikaung, U.S.A.
 Pak, Jinygal, U.S.A.
 Palais, Bob, U.S.A.
 Palais, Richard S., U.S.A.
 Paliouras, John D., U.S.A.
 Palis, Jacob, Brazil
 Palmer, John N., U.S.A.
 Palosaari, Gary C., U.S.A.
 Paluszny, Marco, Venezuela
 Pan, Ting K., U.S.A.
 Pao, Karen, U.S.A.
 Paoli, Madeleine S., U.S.A.
 Papadopol, Petru, U.S.A.
 Papanicolaou, George C., U.S.A.
 Pappas, Peter, U.S.A.
 Pareja, Diego, Colombia
 Parfit, Gary W., U.S.A.
 Park, Sehie, South Korea
 Parker, Phillip E., U.S.A.
 Parker, Thomas H., U.S.A.
 Parker, Willard A., U.S.A.
 Parks, Harold R., U.S.A.

- Parlett, Beresford N., U.S.A.
 Parreau, Michel, France
 Parrott, Mary E., U.S.A.
 Parvizi, Jamshid, Iran
 Pascali, Dan D., U.S.A.
 Pascual-Gainza, Pere, Spain
 Paseman, Gerhard R., U.S.A.
 Passare, Mikael, Sweden
 Pastur, Leonid A., U.S.S.R.
 Pathak, Ram S., India
 Patnaik, Surendra-Nath,
 Algeria
 Patterson, Alan L. T., Scotland
 Paul, Deirdre W., U.S.A.
 Paveri-Fontana, Stefano L.,
 Italy
 Pavlicek, Glenn H., U.S.A.
 Pavone, Marco, U.S.A.
 Pears, Alan R., England
 Pedersen, Gert K., Denmark
 Pedersen, Jean J., U.S.A.
 Pedersen, John F., U.S.A.
 Pedersen, Niels V., Denmark
 Pedrick, George B., U.S.A.
 Pedrosa, Renato L., U.S.A.
 Peetre, Jaak, Sweden
 Pejsachowicz, Jacobo, Italy
 Pekonen, Osmo E. T.,
 Finland
 Pelczar, Andrzej, Poland
 Pellaumail, Jean, France
 Pelles, D. A., U.S.A.
 Pelletier, Donald H., Canada
 Pellikaan, Ruud, Netherlands
 Pelloni, Beatrice, Italy
 Pence, Barbara J., U.S.A.
 Pence, Dennis D., U.S.A.
 Peng, Tsu A., Singapore
 Peretiatkine, Mikhail G.,
 U.S.S.R.
 Perkins, Peter, U.S.A.
 Perović, Miodrag M.,
 Yugoslavia
 Perry, Peter A., U.S.A.
 Pesin, Ya B., U.S.S.R.
 Peters, Galen R., U.S.A.
 Peters, Meinhard H., Federal
 Republic of Germany
 Peterson, Franklin P., U.S.A.
 Peterson, Murray B., U.S.A.
 Peterzil, Kobi A., U.S.A.
 Petro, John, U.S.A.
 Petryshyn, W. V., U.S.A.
 Petty, Clinton M., U.S.A.
 Pfeiffer, Washak F., U.S.A.
 Pfeifer, Richard E., U.S.A.
 Pflugfelder, Hala, U.S.A.
 Phadia, Eswar G., U.S.A.
 Phelps, Andrew R., U.S.A.
 Phelps, Robert R., U.S.A.
 Philippin, Gerard A., Canada
 Philippou, Andreas N., Greece
 Phillips, David L., U.S.A.
 Phillips, N. Christopher, U.S.A.
 Phillips, Ralph S., U.S.A.
 Phillips, Veril L., U.S.A.
 Piccard, D. S., Switzerland
 Piccard, Sophie, Switzerland
 Pickrell, Doug M., U.S.A.
 Picus, Gerald S., U.S.A.
 Pier, Jean-Paul, Luxembourg
 Pierce, Richard S., U.S.A.
 Pillay, Anand, U.S.A.
 Pilz, Günter F., Austria
 Pin, Jean-Eric, France
 Pinchon, Didier M., France
 Pink, Richard, U.S.A.
 Piovani, Luis A., U.S.A.
 Pipher, Jill C., U.S.A.
 Pippenger, Nicholas, U.S.A.
 Pisanski, Tomaž, Yugoslavia
 Pisier, Gilles J., France
 Pitcher, Everett, U.S.A.
 Pitts, David R., U.S.A.
 Pizer, Arnold K., U.S.A.
 Pless, Vera, U.S.A.
 Plymen, Roger J., England
 Poiani, Eileen L., U.S.A.
 Point, Francoise M., Belgium
 Polking, John C., U.S.A.
 Pollack, Daniel, U.S.A.
 Pollack, Richard, U.S.A.
 Pollak, Henry O., U.S.A.
 Pollingher, Adolf, Israel
 Pollington, Andrew D., U.S.A.
 Poluikis, John A., U.S.A.
 Polyakov, Alexander D.,
 U.S.S.R.
 Pomerance, Carl, U.S.A.
 Pong, Glory, Hong Kong
 Ponomarev, Paul, U.S.A.
 Pool, Marsha K., U.S.A.
 Poon, Yat Sun, U.S.A.
 Pope, Clifford O., U.S.A.
 Popov, Vladimir L., U.S.S.R.
 Poritz, Alan B., U.S.A.
 Porte, Jean, France
 Porteous, Hugh L., England
 Porter, Gerald J., U.S.A.
 Porter, James F., U.S.A.
 Poss, Richard L., U.S.A.
 Potthast, John, U.S.A.
 Pour-El, Marian B., U.S.A.
 Power, Stephen C., England
 Prakash, Nirmala, U.S.A.
 Pranger, Walter A., U.S.A.
 PrapaueSSI, Despina Th.,
 U.S.A.
 Prasad, Gopal, India
 Prasad, Phoolan, India
 Prats, Francesc, Spain
 Preston, Jack H., U.S.A.
 Price, David T., U.S.A.
 Price, G. Baley, U.S.A.
 Price, Kenneth H., U.S.A.
 Price, Roderick A., U.S.A.
 Priddy, Stewart B., U.S.A.
 Priestley, William M., U.S.A.
 Prieto, Carlos, Mexico
 Pringle, Thomas H., U.S.A.
 Procesi, Claudio, Italy
 Proctor, Robert A., U.S.A.
 Propp, James G., U.S.A.
 Protter, Murray H., U.S.A.
 Protter, Philip, U.S.A.
 Przytycki, Józef H., Poland
 Pucci, Carlo, Italy
 Pucci, Patrizia, Italy
 Puckette, Stephen E., U.S.A.
 Pugh, Charles C., U.S.A.
 Pugh, Mary C., U.S.A.
 Pukanszky, Layos, U.S.A.
 Pulsinelli, Linda R., U.S.A.
 Puppe, Volker T.,
 Federal Republic of Germany
 Purdy, George B., U.S.A.
 Pustilnik, Seymour W., U.S.A.
 Putnam, Alfred L., U.S.A.
 Putnam, Ian F., U.S.A.
 Puttaswamaiah, Bannikuppe
 M., Canada
 Puttaswamy, T. K., U.S.A.
 Qi Min-you, People's
 Republic of China
 Quan, Luis Jacinto, Guatemala
 Quine, J. R., U.S.A.
 Quinn, Frank, U.S.A.
 Quiroga, Rafael P., U.S.A.
 Rabinowitz, Stanley, U.S.A.
 Racine, Michel L., Canada
 Radjabalipour, Mehdi, Iran
 Raduuskaya, Ami, U.S.A.
 Raghunathan, Madabusi S.,
 India
 Rahimi-Ardabili, Mohammad
 Y., Iran
 Rahman, Mizanur, Canada
 Raimi, Ralph A., U.S.A.
 Rajput, Balram S., U.S.A.
 Rakesh, U.S.A.
 Rakover, Boris D., U.S.A.
 Rallis, Nancy E., U.S.A.
 Ralston, Anthony, U.S.A.
 Ramachandran, Doraiswamy,
 U.S.A.
 Ramanathan, Annamalai,
 U.S.A.
 Ramey, Wade C., U.S.A.
 Ramkissoon, Harold, Trinidad
 Ramsey, Laurence T., U.S.A.

- Rana, David, U.S.A.
 Randall, John D., U.S.A.
 Randić, Milan, U.S.A.
 Range, R. Michael, U.S.A.
 Ranjan, Akhil, India
 Rao, M. M., U.S.A.
 Rao, Veldanda V., Canada
 Rao, Vidhyanath K., U.S.A.
 Raskind, Wayne M., U.S.A.
 Rassias, George M., Greece
 Rassias, Themistocles M., Greece
 Ratcliffe, John G., U.S.A.
 Rathmann, Jürgen, Federal Republic of Germany
 Ratiu, Tudor S., U.S.A.
 Rauch, Gérard, France
 Rausch, Joyce T., U.S.A.
 Rawnsley, John H., England
 Ray, Nigel, England
 Razborov, V. A., U.S.S.R.
 Razenj, Vlado, U.S.A.
 Read, Thomas T., U.S.A.
 Rebman, Kenneth R., U.S.A.
 Redheffer, Raymond M., U.S.A.
 Redlin, Lothar H., U.S.A.
 Redlinger, Reinhard, Federal Republic of Germany
 Reed, Coke S., U.S.A.
 Reed, Diane M., U.S.A.
 Reeken, Michael, Federal Republic of Germany
 Rees, Charles S., U.S.A.
 Rees, Elmer G., Scotland
 Rees, Paul K., U.S.A.
 Reese, Sylvester, U.S.A.
 Reichaw, Meir, Israel
 Reid, K. Brooks, U.S.A.
 Reid, Leslie F., U.S.A.
 Reidhaar-Olson, Lisa, U.S.A.
 Reilly, Ivan L., New Zealand
 Reilly, Robert C., U.S.A.
 Reineck, James F., U.S.A.
 Reinhart, Bruce L., U.S.A.
 Reis, M. Raquel, Portugal
 Reiten, Idun, Norway
 Rejto, Margaret A., U.S.A.
 Rempfer, Robert W., U.S.A.
 Renault, Jean N., France
 Renz, Peter L., U.S.A.
 Repovš, Dušan, Yugoslavia
 Resek, Diane, U.S.A.
 Ressel, Paul, Federal Republic of Germany
 Restrepo, Rodrigo A., Canada
 Reverdy, Jacques, France
 Revuz, Daniel R., France
 Reynolds, Robert R., U.S.A.
 Reza, F. M., Canada
 Rhee, Choon J., U.S.A.
 Rhoades, Billy E., U.S.A.
 Rhodes, John L., U.S.A.
 Ribenboim, Paulo, Canada
 Ribes, Luis, Canada
 Ribet, Kenneth A., U.S.A.
 Rice, Bernard J., U.S.A.
 Richardson Jr., Walter B., U.S.A.
 Richmond, L. Bruce, Canada
 Rickart, Charles E., U.S.A.
 Rickman, Seppo U., Finland
 Rieffel, Eleanor, U.S.A.
 Rieffel, Marc A., U.S.A.
 Rieger, G. J., Federal Republic of Germany
 Riemenschneider, Oswald, Federal Republic of Germany
 Riera, Teresa, Spain
 Riesel, Hans, Sweden
 Rike, Thomas A., U.S.A.
 Riley, Pete E., U.S.A.
 Ringeisen, Richard D., U.S.A.
 Ringseth, Paul F., U.S.A.
 Rinker, Mark W., U.S.A.
 Rinzel, John M., U.S.A.
 Ritter, Gunter, Federal Republic of Germany
 Ritter, Niles D., U.S.A.
 Rivin, Igor, U.S.A.
 Rivlin, Theodore J., U.S.A.
 Roan, Shi-Shyr, China-Taiwan
 Robbiano, Lorenzo, Italy
 Robbins, Neville, U.S.A.
 Roberts, Brooks K., U.S.A.
 Roberts, David P., U.S.A.
 Roberts, Lawrence G., Canada
 Roberts, Leslie G., Canada
 Roberts, Paul C., U.S.A.
 Robertson, Alexander, Australia
 Robertson, G. Neil, U.S.A.
 Robertson, Lewis C., U.S.A.
 Robertson, Wendy J., Australia
 Robin, Guy, France
 Robinson, C. Alan, England
 Robinson, Clark, U.S.A.
 Robinson, Margaret M., U.S.A.
 Robinson, Raphael M., U.S.A.
 Rockwell, Richard D., U.S.A.
 Rodin, Burton, U.S.A.
 Rodrigues, José-Francisco, Portugal
 Rodriguez, Christian, France
 Roe, John, England
 Roesler, Friedrich, Federal Republic of Germany
 Rogawski, Jonathan D., U.S.A.
 Rogers, C. Ambrose, England
 Rogers, James T., U.S.A.
 Rogers Jr., Jack W., U.S.A.
 Röhrle, Gerhard E., Federal Republic of Germany
 Roitman, Moshe, Canada
 Rojo, Javier, U.S.A.
 Romanov, Vladimir, U.S.S.R.
 Rooney, Elaine K., U.S.A.
 Rooney, Paul G., Canada
 Roos, Jan-Erik I., Sweden
 Rose, Henry, South Africa
 Rosenberg, Jonathan M., U.S.A.
 Rosenberg, Steven, U.S.A.
 Rosenberger, Gerhard, Federal Republic of Germany
 Rosenfeld, Melvin, U.S.A.
 Rosenknop, John Z., Federal Republic of Germany
 Rosenlicht, Maxwell, U.S.A.
 Rosenthal, John, U.S.A.
 Rosentrater, C. Ray, U.S.A.
 Rosinski, Jan, U.S.A.
 Ross, Bertram, U.S.A.
 Ross, Kenneth A., U.S.A.
 Ross, Martin, U.S.A.
 Ross, Shepley L., U.S.A.
 Ross II, Shepley L., U.S.A.
 Rossi, Hugo, U.S.A.
 Rossmann, W., Canada
 van Rossum, Marijke, U.S.A.
 Rota, Gian-Carlo, U.S.A.
 Roth, Richard L., U.S.A.
 Rothschild, Bruce, U.S.A.
 Rothschild, Linda P., U.S.A.
 Rottenberg, Reuven R., Israel
 Rovnyak, James, U.S.A.
 Rowan, William H., U.S.A.
 Rowley, C. A., England
 Roy, Lucien R., U.S.A.
 Roy, Nina M., U.S.A.
 Royden, Halsey L., U.S.A.
 Rozanov, Youri A., U.S.S.R.
 Ruan, Zhong-Jin, U.S.A.
 Rubenfeld, Lester A., U.S.A.
 Ruberman, Daniel, U.S.A.
 Rubin, Karl, U.S.A.
 Rubinstein, Joachim H., Australia
 Rudin, Mary E., U.S.A.
 Rudin, Walter, U.S.A.
 Ruelle, David P., France
 Ruh, Ernst A., U.S.A.
 Rumbos, Adolfo J., U.S.A.
 Rumely, Robert S., U.S.A.
 Rung, Donald C., U.S.A.
 Rush, David E., U.S.A.
 Ruskai, M. Beth, U.S.A.
 Russek, Andrzej, U.S.A.
 Russell, Dennis C., Canada
 Russell, Karl P., Canada
 Russo, Bernard, U.S.A.

- Russo, David A., U.S.A.
 Russo, Paula A., U.S.A.
 Rutter, John W., England
 Ryan, Patrick J., Canada
 Rychlik, Marek R., U.S.A.
 Ryeng, Torbjörn E., Sweden
 Ryff, John V., U.S.A.
 Rykken, Charles J., U.S.A.
 Rylands, Leanne J., Australia
 Rzedowski, Martha, U.S.A.
 Saad, Gerhard, Federal
 Republic of Germany
 Saba, Farrokh, U.S.A.
 Sachs, Eliza M., U.S.A.
 Sad, Paulo R. G., Brazil
 Safont, Carmen, U.S.A.
 Sagan, Bruce E., U.S.A.
 Sagie, Arthur A., U.S.A.
 Sahab, Salem A., Saudi
 Arabia
 Saidi, Samira, U.S.A.
 Saito, Mutsumi, U.S.A.
 Saitoh, Saburou, Japan
 Sakai, Akira, Japan
 Sakai, Shôichirô, Japan
 Sakarovitch, Jacques, France
 Salamanco Riba, Susana A.,
 U.S.A.
 Salehi, Ebrahim, U.S.A.
 Saletan, Eugene J., U.S.A.
 Salisbury, Thomas S., Canada
 Sally, Paul J., U.S.A.
 Salvadori, Anna, Italy
 Salzer, Herbert E., U.S.A.
 Samelson, Hans, U.S.A.
 Sampson, Richard W., U.S.A.
 Sanatani, Sunoy, Canada
 Sandar, Bradford L., U.S.A.
 Sands, Arthur D., Scotland
 Sands, Jonathan W., U.S.A.
 Sangare, Daouda, Ivory Coast
 Sanker, David V., U.S.A.
 Sarafyan, Diran, U.S.A.
 Sarason, Donald E., U.S.A.
 Sarnak, Peter C., U.S.A.
 Sasano, Kazuhiro, Japan
 Sastri, Chelluri C. A., Canada
 Sastry, Darbha Ramakrishna,
 India
 Satake, Ichiro, Japan
 Sather, Jerome, U.S.A.
 Sato, Atsushi, Japan
 Sato, Hajime, Japan
 Sato, Masahisa, Japan
 Sauer, Niko, South Africa
 Saunders, Frank W., U.S.A.
 Saut, Jean-Claude, France
 Sawada, Ken, Japan
 Sawka, John M., U.S.A.
 Sawyer, Eric T., Canada
 Saxena, Subhash C., U.S.A.
 Saxl, Jan, England
 Saxton, Ralph A., U.S.A.
 Sayrafiezadeh, Mahmoud, Iran
 Scanlan, Michael J., U.S.A.
 Scaramuzzi, Roberto, U.S.A.
 Schacher, Murray M., U.S.A.
 Schafer, Alice T., U.S.A.
 Schaffer, Richard D., U.S.A.
 Schäffer, Alejandro A., U.S.A.
 Schäffer, Juan J., U.S.A.
 Schaffer, Matthew J., U.S.A.
 Schaps, Mary E., Israel
 Scharlemann, Martin G., U.S.A.
 Schattschneider, Doris J.,
 U.S.A.
 Schechter, Samuel, U.S.A.
 Scheepers, Marion, U.S.A.
 Scheffer, Vladimir, U.S.A.
 Schep, Anton R., U.S.A.
 Schilling, Kenneth E., U.S.A.
 Schirmer, Helga H., Canada
 Schirokauer, Oliver A., U.S.A.
 Schlachter, Sandra A., U.S.A.
 Schlichting, Günter H.,
 Federal Republic of Germany
 Schmeelk, John, U.S.A.
 Schmeichel, Edward F., U.S.A.
 Schmerl, James H., U.S.A.
 Schmickler-Hirzebruch, Ulrike,
 Federal Republic of Germany
 Schmid, Peter P., Federal
 Republic of Germany
 Schmid, Roland J., Switzerland
 Schmid, Rudolf, U.S.A.
 Schmid, Wilfried, U.S.A.
 Schmidt, Dieter F. H.,
 Federal Republic of Germany
 Schmidt, Siegbert, Federal
 Republic of Germany
 Schmitt, Frederick G., U.S.A.
 Schmitz, Udo, Federal
 Republic of Germany
 Schnare, Paul S., U.S.A.
 Schochet, Claude, U.S.A.
 Schoen, Alan H., U.S.A.
 Schoen, Richard M., U.S.A.
 Schoenfeld, Lowell, U.S.A.
 Schoisengeier, Johannes,
 Austria
 Scholl, Anthony J., England
 Schonbek, Maria E., U.S.A.
 Schönhage, Arnold, Federal
 Republic of Germany
 Schoof, Rene, U.S.A.
 Schori, Richard M., U.S.A.
 Schreiber, Bertram M., U.S.A.
 Schrijver, Alexander,
 Netherlands
 Schroeder, Viktor M., Federal
 Republic of Germany
 Schulte, Egon, Federal
 Republic of Germany
 Schultz, Henry J., U.S.A.
 Schultz, Reinhard E., U.S.A.
 Schulz, Christoph, Federal
 Republic of Germany
 Schulze-Pillot, Rainer, Federal
 Republic of Germany
 Schumer, Peter, U.S.A.
 Schuster, Hans W., Federal
 Republic of Germany
 Schuster, Seymour, U.S.A.
 Schwab, Margaret R., U.S.A.
 Schwartz, Helga, U.S.A.
 Schwartz, Jacob T., U.S.A.
 Schwartz, Jean-Marie, France
 Schwarz, Binyamin, Israel
 Schwarz, Gerald W., U.S.A.
 Schwarz, Wolfgang, Federal
 Republic of Germany
 Schweitzer, Eric, U.S.A.
 Schweitzer, Paul A., Brazil
 Schweizer, David, U.S.A.
 Scott, G. Peter, England
 Scourfield, Eira J., England
 Scowcroft, Philip H., U.S.A.
 Seddighi, Karim, Iran
 Seding, Harry, U.S.A.
 Seebach, J. Arthur, U.S.A.
 Segal, Irving E., U.S.A.
 Seidenberg, Abraham, U.S.A.
 Seitz, Gary M., U.S.A.
 Seiya, Sasao, Japan
 Sekimoto, Toshihiko, Japan
 Selfridge, John L., U.S.A.
 Selick, Paul S., Canada
 Seligman, George B., U.S.A.
 Sell, George R., U.S.A.
 Semadeni, Zbigniew W.,
 Poland
 Sempi, Carlo E., Italy
 Sendov, Blagovest, Bulgaria
 Sepikas, John P., U.S.A.
 Seppälä, Mika, Federal
 Republic of Germany
 Seppala-Holtzman, David N.,
 U.S.A.
 Series, Caroline, England
 Serrano-Garcia, Fernando,
 U.S.A.
 Serre, Jean-Pierre, France
 Sesay, Mohamed W. I., U.S.A.
 Seshadri, Manavasi S.,
 Federal Republic of Germany
 Sestini, Sally D., U.S.A.
 Seydi, Hamet, Senegal
 Seymour, Paul D., U.S.A.
 Sha, Ji-Ping, U.S.A.
 Shababi, Nastaran, Iran
 Shadman, Dariush, Iran
 Shadwick, William F., Canada

- Shafer, Robert E., U.S.A.
 Shaffer, Dorothy B., U.S.A.
 Shahriari, Shahriar, U.S.A.
 Shahshahani, Mehrdad M., U.S.A.
 Shahshahani, Siavash M., Iran
 Shalen, Peter B., U.S.A.
 de Shalit, Ehud, U.S.A.
 Shamir, Adi, Israel
 Shamma, Amal K., U.S.A.
 Shamma, Shawky E., U.S.A.
 Shanks, Daniel, U.S.A.
 Shapiro, Dawn M., U.S.A.
 Shapiro, Jesse M., Israel
 Shapiro, Louis W., U.S.A.
 Shapiro, Victor L., U.S.A.
 Sharp, Henry, U.S.A.
 Shatz, Stephen S., U.S.A.
 Shaw, Mei-Chi, U.S.A.
 Shearer, James W., U.S.A.
 Shekoury, Raymond N., Iraq
 Shelah, Saharon, Israel
 Shelley, Michael J., U.S.A.
 Sherman, Clayton C., U.S.A.
 Sheu, Shey Shiung, China-Taiwan
 Shick, Paul L., U.S.A.
 Shiffman, Max, U.S.A.
 Shifrin, Theodore, U.S.A.
 Shim, Jae Ung, South Korea
 Shimada, Nobuo, Japan
 Shimamoto, Don H., U.S.A.
 Shin, Kil-Ho, Japan
 Shishikura, Mitsuhiro, Japan
 Shiu, Peter, England
 Shiue, Peter J., U.S.A.
 Shokranian, Salahoddin, Brazil
 Shokurov, V. V., U.S.S.R.
 Shor, Peter W., U.S.A.
 Shub, Michael, U.S.A.
 Shubin, Carol A., U.S.A.
 Shukla, Pradeep K., U.S.A.
 Shulman, Bonnie J., U.S.A.
 Shum, Kar-Ping, Hong Kong
 Siafarikas, Panagiotis, Greece
 Sibirski, Constantin S., U.S.S.R.
 Sibley, Thomas Q., U.S.A.
 Šidák, Zbyněk, Czechoslovakia
 Siddiqi, Jamil A., Canada
 Siebenmann, Laurent, France
 Sieburg, Hans B., U.S.A.
 Siegel, David, Canada
 Siegfried, Edward, U.S.A.
 Siemons, Johannes, England
 Siersma, Dirk, Netherlands
 Sigman, Karl, U.S.A.
 Sigman, Thomas L., U.S.A.
 Sigurdsson, Ragnar, Iceland
 Silberstein, Esther, U.S.A.
 Silva, Cesar E., U.S.A.
 Silver, Ben, U.S.A.
 Silver, Jack H., U.S.A.
 Silverberg, Alice, U.S.A.
 Silverman, Robert D., U.S.A.
 Silvia, Evelyn M., U.S.A.
 Simon, Arthur B., U.S.A.
 Simons, Gordon E., Canada
 Simons, James H., U.S.A.
 Simpson, Henry C., U.S.A.
 Simpson Stephen G., U.S.A.
 Sims, Benjamin T., U.S.A.
 Sims, Charles C., U.S.A.
 Simutis, Jean S., U.S.A.
 Sine, Robert C., U.S.A.
 Singer, I. M., U.S.A.
 Singer, Michael F., U.S.A.
 Singh, Dinesh, India
 Singh, Udit N., India
 Singmaster, David, England
 Sipser, Michael, U.S.A.
 Siu Man-Keung, Hong Kong
 Siu, Yum-Tong, U.S.A.
 Sjölin, Per B., Sweden
 Skalba, Justine, U.S.A.
 Sklar, David, U.S.A.
 Skora, Richard K., U.S.A.
 Skorohod, A. V., U.S.S.R.
 Skoug, David L., U.S.A.
 Skrien, Dale J., U.S.A.
 Skrypnik, Igor V., U.S.S.R.
 Slack, Michael D., U.S.A.
 Slaman, Theodore A., U.S.A.
 Slater, Harold, U.S.A.
 Sloss, James M., U.S.A.
 Sloughter, Daniel C., U.S.A.
 Sloyan, Sister Stephanie, U.S.A.
 Smale, Nat, U.S.A.
 Smale, Steve, U.S.A.
 Smalø, Sverre O., Norway
 Smart, James R., U.S.A.
 Smillie, John, U.S.A.
 Smith, Alex J., U.S.A.
 Smith, David A., U.S.A.
 Smith, Doris A., U.S.A.
 Smith, J. C., U.S.A.
 Smith, John M., U.S.A.
 Smith, Marianne F., U.S.A.
 Smith, Mark H., U.S.A.
 Smith, Martha K., U.S.A.
 Smith, Murray H., New Zealand
 Smith, Richard A., U.S.A.
 Smith, Roy C., U.S.A.
 Smith, Stuart P., U.S.A.
 Smith, Tara L., U.S.A.
 Smith, Wayne E., U.S.A.
 Smith, Wayne S., U.S.A.
 Smithson, R. E., U.S.A.
 Smoke, W. H., U.S.A.
 Smolensky, Roman L., U.S.A.
 Smolinsky, Lawrence J., U.S.A.
 Smorynski, Craig A., U.S.A.
 Snapper, Ernst, U.S.A.
 Snow, Donald R., U.S.A.
 So, Joseph W.-H., U.S.A.
 Soare, Robert I., U.S.A.
 Soedirman, Massy, France
 Soemartojo, Noeniek, Indonesia
 Sogge, Christopher D., U.S.A.
 Solarin, Adewale R. T., Nigeria
 Soleng, Ragnar, Norway
 Solomon, David R., U.S.A.
 Solonnikov, Vsevolod A., U.S.S.R.
 Solovay, Robert M., U.S.A.
 Somer, Lawrence E., U.S.A.
 Sommer, Rick D., U.S.A.
 Soni, Kusum K., U.S.A.
 Soni, Raj Pal, U.S.A.
 Sonn, Jack, Israel
 Sons, Linda R., U.S.A.
 Sontag, Alexia, U.S.A.
 Soon, Frederick H., U.S.A.
 Sorin, Sylvain P., France
 Soto-Andrade, Jorge, Chile
 Sottile, Frank, England
 Soulé, Christophe J., France
 Soules, George W., U.S.A.
 Sourour, A. R., Canada
 Spake, Reuben, U.S.A.
 Spalt, Detlef, Federal Republic of Germany
 Spanier, Edwin H., U.S.A.
 Spanier, Fred, U.S.A.
 Specht, Edward J., U.S.A.
 Speck, Frank-Olme, Federal Republic of Germany
 Spector, Robert M., U.S.A.
 Speed, Cynthis A., U.S.A.
 Speed, Terence P., Australia
 Speer, Eugene R., U.S.A.
 Speh, Birgit E. M., U.S.A.
 Spencer, Thomas, U.S.A.
 Spielberg, Jack, U.S.A.
 Spielberg, Stephen E., U.S.A.
 Spinadel, Vera W. de, Argentina
 Spinelli, Giancarlo, Italy
 Spira, Michel, Brazil
 Spiro, Claudia A., U.S.A.
 Spring, David, Canada
 Springer, Arthur, U.S.A.
 Spurr, Michael J., U.S.A.
 Srinivasan, T. P., U.S.A.
 Srivastava, Jaya, U.S.A.

- Stackelberg, Cora E., U.S.A.
 Stackelberg, Olaf P., U.S.A.
 Stahl, Neil, U.S.A.
 Stakgold, Ivar, U.S.A.
 Stallings, John R., U.S.A.
 Stampfli, Joseph G., U.S.A.
 Stanford, Theron, U.S.A.
 Stanković, Bogoljub,
 Yugoslavia
 Stanley, Maurice C., U.S.A.
 Stanton, Dennis W., U.S.A.
 Stanton, Nancy K., U.S.A.
 Starbird, Michael, U.S.A.
 Starbird, Thomas W., U.S.A.
 Stark, Christopher W., U.S.A.
 Stark, Harold M., U.S.A.
 Starr, Norton, U.S.A.
 Steel, John R., U.S.A.
 Steen, Lynn A., U.S.A.
 Steenbrink, Joseph,
 Netherlands
 Stefánsson, Jón R., Iceland
 Stegenga, David A., U.S.A.
 Steger, Tim J., U.S.A.
 Stein, Elias M., U.S.A.
 Stein, Harvey J., U.S.A.
 Stein, Michael R., U.S.A.
 Steinberg, Leon, U.S.A.
 Steinberg, Maria, U.S.A.
 Steinberg, Robert, U.S.A.
 Steinert, Leon A., U.S.A.
 Steinhorn, Charles I., U.S.A.
 Stembridge, John R., U.S.A.
 Stephens, Clarence F., U.S.A.
 Stephens, Harriette J., U.S.A.
 Stephenson, Kenneth, U.S.A.
 Stern, Lynnell E., U.S.A.
 Stern, Ronald J., U.S.A.
 Sterrett, Andrew, U.S.A.
 Stevens, Damon R., U.S.A.
 Stevens, Glenn H., U.S.A.
 Stob, Michael J., U.S.A.
 Stojanov, Luchezar N.,
 Bulgaria
 Stokes, Christine B., U.S.A.
 Stokes, Russell A., U.S.A.
 Stoltzfus, Neal W., U.S.A.
 Stone, Charles J., U.S.A.
 Stone, Marshall H., U.S.A.
 Stone, Michael G., Canada
 Stone, William D., U.S.A.
 Størmer, Erling, Norway
 Stout, Edgar L., U.S.A.
 Stowasser, Roland J., Federal
 Republic of Germany
 Straffin, Philip D., U.S.A.
 Straight, H. Joseph, U.S.A.
 Strang, Gilbert, U.S.A.
 Strano, Rosario, Italy
 Strassen, Volker, Switzerland
 Stratton, Beverly J., U.S.A.
 Straubing, Howard, U.S.A.
 Straume, Eldar J., Norway
 Strauss, Dona A., England
 Stray, Arne, Norway
 Strickland, Elisabetta, Italy
 Strooker, Jan R., Netherlands
 Stuart, Jeff L., U.S.A.
 Sturmfels, Bernd, U.S.A.
 Stys, Tadeusz, Poland
 Su, Weiyl, U.S.A.
 Subrao, Doré R., U.S.A.
 Sullivan, Dennis, U.S.A.
 Sumihiro, Hideyasu, U.S.A.
 Summerville, Richard M.,
 U.S.A.
 Sun, Hugo S.-H., U.S.A.
 Sun, Shunhua, People's
 Republic of China
 Sun, Tze-Chien, U.S.A.
 Sund, Terje, Norway
 Sundaresan, K., U.S.A.
 Sundberg, Carl, U.S.A.
 Sunder, Viakalathur S., India
 Sung, Chen-Han, U.S.A.
 Sunley, Judith S., U.S.A.
 Suri, Manil, U.S.A.
 Surin, Aline, France
 Suslin, Andrei A., U.S.S.R.
 Sussmann, Hector J., U.S.A.
 Suwa, Tatsuo, Japan
 Suzuki, Masashi, Japan
 Suzuki, Michio, U.S.A.
 Svensson, Erik, Sweden
 Sverdløve, Ronald, U.S.A.
 Swaminathan, Srinivasa,
 Canada
 Sward, Marcia P., U.S.A.
 Swart, Johan, South Africa
 Symonds, Peter S., U.S.A.
 Synowicz, John A., U.S.A.
 Szabó, Zoltán I., Hungary
 Szarek, Stanislaw J., U.S.A.
 Szenthe, János, Hungary
 Szeto, Mabel, U.S.A.
 Szőke, Róbert, U.S.A.
 Szpirglas, Viviane, France
 Szpirglass, Jacques, France
 Taam, Choy-Tak, U.S.A.
 Taft, Earl J., U.S.A.
 Tagawa, Shojiro, Japan
 Taguchi, Yoshiko, Japan
 Takahashi, Shuichi, Japan
 Takeno, Hyoichiro, Japan
 Takens, Floris, Netherlands
 Takesaki, Masamichi, U.S.A.
 Takizawa, Seiji, Japan
 Talli, Mahmoud R., Iran
 Tamaki, Robert K., U.S.A.
 Tamari, Dov, U.S.A.
 Tamburini, Maria Clara, Italy
 Tamm, Martin C. A., Sweden
 Tamura, Takayuki, U.S.A.
 Tan, Henry K., U.S.A.
 Tanaka, Kazuyuki, U.S.A.
 Tanbay, Betül, U.S.A.
 Tang, Ping Tak P., U.S.A.
 Tang, Tai-Man, U.S.A.
 Tangerman, Folkert M., U.S.A.
 Tangora, Martin C., U.S.A.
 Tannenbaum, Peter, U.S.A.
 Tardos, Eva, U.S.A.
 Tashiro, Yoshiaki, Japan
 Tata, Mahabano N., Iran
 Tatevossian, Leon H., U.S.A.
 Tatsuoka, Kay S., U.S.A.
 Tattersall, James J., U.S.A.
 Taub, Abraham H., U.S.A.
 Taubes, Clifford H., U.S.A.
 Taylor, Angus E., U.S.A.
 Taylor, Francis B., U.S.A.
 Taylor, Gerald D., U.S.A.
 Taylor, John, Jamaica
 Taylor, Laurence R., U.S.A.
 Tchere, Seka, France
 Tebbs, Richard R., U.S.A.
 Tee, Garry J., New Zealand
 Teitelbaum, Jeremy T., U.S.A.
 Teleman, Nicholas, U.S.A.
 Teles, Elizabeth J., U.S.A.
 Teller, Joel H., U.S.A.
 Temple, John B., U.S.A.
 Tenorio, Luis-Francisco, U.S.A.
 Terada, Itaru, Japan
 Terlinden, D. M., U.S.A.
 Terng, Chuu-Lian, U.S.A.
 Terry, Raymond D., U.S.A.
 Teterin, Alexandr G., U.S.S.R.
 Tews, Melvin C., U.S.A.
 Textorius, Björn O. B.,
 Sweden
 Tezuka, Michishige, Japan
 Thakur, Dinesh S., U.S.A.
 Thaler, Alvin I., U.S.A.
 Thayer, F. Javier, U.S.A.
 Thelen, Brian J., U.S.A.
 Therien, Denis, Canada
 Thi, Dao trong, Vietnam
 Thickstun, Thomas L., U.S.A.
 Thiel, Linda C., U.S.A.
 Thomas, Ben, U.S.A.
 Thomas, James H., U.S.A.
 Thomas, Jay F., U.S.A.
 Thomas, Ken E., U.S.A.
 Thomas, Marc P., U.S.A.
 Thomas, P. Emery, U.S.A.
 Thomas, Pascal J., U.S.A.
 Thomas, Robert S. D., Canada
 Thomason, Robert W., U.S.A.
 Thompson, George G., U.S.A.

- Thompson, Thomas M., U.S.A.
 Thomsen, Momme J., Federal
 Republic of Germany
 Thomson, Brian S., Canada
 Thorne, Ann T., U.S.A.
 Thorne, John F., U.S.A.
 Thorup, Anders, Denmark
 Throop, T., U.S.A.
 Tian Jinghuang, People's
 Republic of China
 Tiahrt, Chris A., U.S.A.
 Tichy, Robert F., Austria
 Timmerscheidt, Klaus, Federal
 Republic of Germany
 Timourian, James G., Canada
 Tingley, Daryl W., Canada
 Tippet, Gary R., U.S.A.
 Tkatchenko, Vladimir I.,
 U.S.S.R.
 Togari, Yoshio, Japan
 Tôgô, Shigeaki, Japan
 Tokuyama, Takeshi, Japan
 Toledo, Domingo, U.S.A.
 Tollis, Theodore, U.S.A.
 Tolosa, Juan J., Venezuela
 Tolsted, Elmer B., U.S.A.
 Tom Wan Yan-Heng, Hong
 Kong
 Tomaru, Tadashi, Japan
 Tomei, Carlos, Brazil
 Tomida, Masamichi, Japan
 Tominaga, Akira, Japan
 Tomiyama, Jun, Japan
 Tomter, Per, Norway
 Tondeur, Philippe, U.S.A.
 Tondra, Richard J., U.S.A.
 Tong, Po, U.S.A.
 Tong, Yue Lin L., U.S.A.
 Tonga, Marcel, Canada
 Topiwala, Pankaj N., U.S.A.
 Toppila, Sakari O., Finland
 Topsøe, Flemming, Denmark
 Torretto, Roseanna F., U.S.A.
 Totaro, Burt J., U.S.A.
 Touré, Saliou, Ivory Coast
 Towber, Jacob, U.S.A.
 Towers, David A., England
 Townsend, Ralph N., U.S.A.
 Toyoda, Masanori, Japan
 Tracy, Craig A., U.S.A.
 Traczyk, Paweł, Poland
 Traub, Joseph F., U.S.A.
 Trautmann, Guenther, Federal
 Republic of Germany
 Treiman, Jay, U.S.A.
 Tromba, Anthony J., U.S.A.
 Trombi, Pete C., U.S.A.
 Tropp, Henry S., U.S.A.
 Trotman, David J. A., France
 Troyanov, Marc, Switzerland
 Troyer, Stephanie F., U.S.A.
 Troyer, Todd W., U.S.A.
 Trudinger, Neil S., Australia
 Tsai, Chester, U.S.A.
 Tsatsanis, Peter S., Canada
 Tschantz, Steven T., U.S.A.
 Tsou, Sheung Tsun, England
 Tsuboi, Akito, Japan
 Tsuboi, Takashi, Japan
 Tsuchiya, Takuya, Japan
 Tsuda, Teruko, Japan
 Tsuji, Hajime, Japan
 Tsukada, Haruo, Japan
 Tsunoda, Shuichiro, U.S.A.
 Tu, Loring W., U.S.A.
 Tucker, John R., U.S.A.
 Tucker, Thomas W., U.S.A.
 Tukia, Pekka, Finland
 Turgeon, Jean M., Canada
 Turner, Walter W., U.S.A.
 Turpin, Philippe, France
 Turquette, Atwell R., U.S.A.
 Turyn, Richard J., U.S.A.
 Tyler, Douglas B., U.S.A.
 Tysk, Johan, U.S.A.
 Ue, Masaaki, Japan
 Uehara, Hiroshi, U.S.A.
 Ueno, Tadashi, Japan
 Uherka, David J., U.S.A.
 Uhlenbeck, Karen K., U.S.A.
 Ulbrich, Karl-Heinz, Federal
 Republic of Germany
 Ullman, Daniel, U.S.A.
 Ulmer, Douglas L., U.S.A.
 Ulucay, Cengiz, Turkey
 Umeda, Takato, U.S.A.
 Umehara, Masaaki, Japan
 Underwood, Douglas H.,
 U.S.A.
 Ungar, Jerry, U.S.A.
 Ungar, Peter, U.S.A.
 Upadhyayula, Satyanarayana
 V., U.S.A.
 Upatisinga, Vis, U.S.A.
 Upmeier, Harald, U.S.A.
 Urabe, Tohsuke, Japan
 Ural' tseva, N. N., U.S.S.R.
 Urban, Carol H., U.S.A.
 Urbanik, Kazimierz, Poland
 Uy, Nguyen X., U.S.A.
 Uzawa, Tohru, U.S.A.
 Vaillancourt, Rémi, Canada
 Vakilian, Ramin, U.S.A.
 Valente, Ken, U.S.A.
 Valette, Alain J., Belgium
 Valiant, Leslie G., U.S.A.
 Vamos, Peter, England
 Van Arkel, Nicolaas A.,
 Nigeria
 Van Brundt, Hendrik, U.S.A.
 Van Daele, Alfons, Belgium
 Van de Wetering, R. Lee,
 U.S.A.
 Van Den Berg, Jacob,
 Netherlands
 Van Der Kallen, Wilberd L.,
 Netherlands
 Van Eldik, Peter, South Africa
 Van Geel, Jan M.-H., Belgium
 Van Lint, Jacobus H.,
 Netherlands
 Van Osdol, Donovan H., U.S.A.
 Van, Tran Duc, Vietnam
 VanBuskirk, Jim, U.S.A.
 Vanhecke, Lieven, Belgium
 Vannini, Walter, Australia
 Varadarajan, V. S., U.S.A.
 Vargas, Jose M., U.S.A.
 Vasco, Carlos E., Colombia
 Vaughn, James M., Vaughn,
 U.S.A.
 Vaught, Robert L., U.S.A.
 Vazirani, Umesh V., U.S.A.
 Vazirani, Vijay V., U.S.A.
 Velling, John A., U.S.A.
 Venema, Gerard A., U.S.A.
 Ventura, Belisario A., U.S.A.
 Venzke, Paul B., U.S.A.
 Verdonk, Brigitte, Belgium
 Verma, Krishnanand, U.S.A.
 Verma, Sadanand, U.S.A.
 Vernon, Micheal H., U.S.A.
 Verona, Andrei, U.S.A.
 Verrios, Katherine G. D.,
 Greece
 Verwoert, J. W. G., Netherlands
 Vessey, Theodore A., U.S.A.
 Vicknair, J. Paul, U.S.A.
 Videla, Carlos R., U.S.A.
 Viehweg, Eckart E., Federal
 Republic of Germany
 Vila-Freyer, Ricardo F., U.S.A.
 Villani, Vinicio, Italy
 Villar, Luzviminda V.,
 Philippines
 Villari, Gabriele, Italy
 Vilms, Jaak, U.S.A.
 Vince, Andrew, U.S.A.
 Vincent, Georges A.,
 Switzerland
 Vincze, Istvan, Hungary
 Vineberg, Susan N., U.S.A.
 Vinti, Calogero, Italy
 Vitulli, Marie A., U.S.A.
 Vlach, Alan D., U.S.A.
 Vladimirov, Vasilii S., U.S.S.R.
 Vo Khac, Khoan, France
 Vogan, David A., U.S.A.
 Vogel, Annie, France
 Vogel, Pierre, France
 Vogtmann, Karen L., U.S.A.

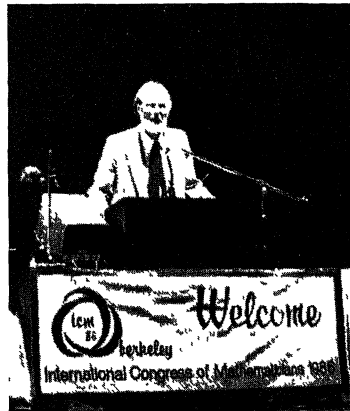
- Voiculescu, Dan, Romania
 Vojta, Paul A., U.S.A.
 Vought, Eldon J., U.S.A.
 Voytuk, James A., U.S.A.
 Vrem, Richard C., U.S.A.
 Vukomanović, Djordje V., Yugoslavia
 Vuorinen, Matti, Finland
 Wachs, Michelle L., U.S.A.
 Wada, Hidekazu, Japan
 Wade, William R., U.S.A.
 Wadleigh, Frank R., U.S.A.
 Wagner, Catherine M., U.S.A.
 Wagoner, Jack, U.S.A.
 Wagreich, Philip, U.S.A.
 Wagstaff, Samuel S., U.S.A.
 Wahl, Jonathan M., U.S.A.
 Waits, Bert K., U.S.A.
 Wald, Linda A., U.S.A.
 Walker, James D., U.S.A.
 Walker, Kevin M., U.S.A.
 Walker, LeRoy H., U.S.A.
 Walker, Peter L., England
 Walkoe, Iver A., U.S.A.
 Wall, Gordon E., Australia
 Walsh, John J., U.S.A.
 Walsh, Timothy R., Canada
 Walter, Gilbert G., U.S.A.
 Walter, Wolfgang L., Federal Republic of Germany
 Walters, Peter, England
 Walters, Randall K., U.S.A.
 Wan, Yieh-Hei, U.S.A.
 Wan, Zhe-Xian, People's Republic of China
 Wang, Ai Nung, U.S.A.
 Wang, Kwang-Shang, U.S.A.
 Wang, Qiming, People's Republic of China
 Wang Rouhuai, People's Republic of China
 Wang, Stuart S.-S., U.S.A.
 Wang, Tseng-Chan, U.S.A.
 Wang, Xiaolu, U.S.A.
 Wang, Yar-Yi, U.S.A.
 Warchall, Henry, U.S.A.
 Ward, James E., U.S.A.
 Ware, Buck, U.S.A.
 Warren, Bette L., U.S.A.
 Warren, Richard H., U.S.A.
 Washburn, Sherwood, U.S.A.
 Washington, Lawrence C., U.S.A.
 Wasilkowski, Grzegorz W., U.S.A.
 Wassermann, Antony J., England
 Watanabe, Kimio, Japan
 Watanabe, Masayuki, Japan
 Watanabe, Toshihiro, Japan
 Waterman, Daniel, U.S.A.
 Waters, Steven R., U.S.A.
 Wathen, Andrew, England
 Watkins, Joseph C., U.S.A.
 Watson, Martha F., U.S.A.
 Watson, Saleem H., U.S.A.
 Weaver, John C., U.S.A.
 Webb, Peter J., England
 Webber, David B., Scotland
 Wedel, Arnold M., U.S.A.
 Wegner, Gerd, Federal Republic of Germany
 Wei, S. Walter, U.S.A.
 Weibel, Charles A., U.S.A.
 Weill, Georges G., France
 Weinberg, David A., U.S.A.
 Weinberger, Hans F., U.S.A.
 Weiner, Joel L., U.S.A.
 Weinerfelt, Per A., Sweden
 Weinstein, Alan D., U.S.A.
 Weinstein, Lenard M., U.S.A.
 Weinstein, Michael I., U.S.A.
 Weintraub, Steven H., U.S.A.
 Weisfeld, Morris, U.S.A.
 Weisinger, James R., U.S.A.
 Weiss, Gary L., U.S.A.
 Weiss, Norman J., U.S.A.
 Weldon, Robert J., U.S.A.
 Wells, Benjamin, U.S.A.
 Wells, R.O., U.S.A.
 Welty, Christopher S., U.S.A.
 Wen Guo-chun, Federal Republic of Germany
 Wen Shih-liang, U.S.A.
 Wentz, Henry C., U.S.A.
 Werdier, Dieter, Federal Republic of Germany
 Wermer, John, U.S.A.
 Westcott, Alan J., U.S.A.
 Westman, Joel J., U.S.A.
 Westphal-Schmidt, Ursula, Federal Republic of Germany
 Westwood, Derek J., U.S.A.
 Weyland, Nicholas J., U.S.A.
 Wheeler, David H., Canada
 White, Alvin, U.S.A.
 White, Arthur T., U.S.A.
 White, Benjamin S., U.S.A.
 White, Brian C., U.S.A.
 White, Dennis E., U.S.A.
 White, Neil L., U.S.A.
 White, Samuel P., U.S.A.
 White, Tad P., U.S.A.
 Whitman, Andrew P., Brazil
 Whitney, Hassler, U.S.A.
 Whitney, Stephen, Canada
 Whyburn, Lucille E., U.S.A.
 Wickerhauser, Mladen V., U.S.A.
 Widom, Harold, U.S.A.
 Wiener, Howard W., U.S.A.
 Wiener, Matthew P., U.S.A.
 Wiid, Frans G., South Africa
 Wilkens, David L., England
 Wilkerson, Clarence W., U.S.A.
 Wilkie, Alec J., England
 Willcox, Alfred B., U.S.A.
 Wille, Friedrich, Federal Republic of Germany
 Williams, Dana P., U.S.A.
 Williams, Daniel A., U.S.A.
 Williams, Hugh C., Canada
 Williams, Janet, U.S.A.
 Williams, Luis A., Nicaragua
 Williams, Mark, U.S.A.
 Williams, Robert F., U.S.A.
 Williamson, Robert E., U.S.A.
 Wilson, John H., U.S.A.
 Wilson, John S., England
 Wilson, Leslie C., U.S.A.
 Wilson, Pelham M. H., England
 Wilson, Robert L., U.S.A.
 Wilson, Robert R., U.S.A.
 Wingren, Peter A. F., Sweden
 Winkelkemper, Horst E., U.S.A.
 Wirszup, Izaak, U.S.A.
 Wisniewski, Helena S., U.S.A.
 Witte, Dave, U.S.A.
 Witten, Edward, U.S.A.
 Wogen, Warren R., U.S.A.
 Wojciechowski, Krzysztof P., Poland
 Wolf, Edwin M., U.S.A.
 Wolf, Joseph A., U.S.A.
 Wolf, Robert S., U.S.A.
 Wolfmann, Jacques, France
 Wolfram, Suzanne, U.S.A.
 Wolfskill, John C., U.S.A.
 Wolfson, Jon G., U.S.A.
 Wolpert, Scott A., U.S.A.
 Wong, Chi Song, Canada
 Wong, M. W., Canada
 Wong, Yan-Loi, U.S.A.
 Wongkew, Richard A., U.S.A.
 Woo, Norman T., U.S.A.
 Wood, Carol S., U.S.A.
 Wood, Christopher M., England
 Wood, Dick A., U.S.A.
 Wood, Graham R., New Zealand
 Wood, Jay A., U.S.A.
 Woodin, W. Hugh, U.S.A.
 Woods, Alan, U.S.A.
 Woods, Alan C., U.S.A.
 Woods, Dale, U.S.A.
 Woods, E. James, Canada
 Woodside, Robert M., U.S.A.
 Woolf, Thomas, U.S.A.

Woolf, William B., U.S.A.
 Wooster, Kenneth, U.S.A.
 Youafu, Jean Kamga,
 Cameroon
 Wouk, Arthur, U.S.A.
 Woźniakowski, Henryk, U.S.A.
 Wright, David G., U.S.A.
 Wright, Jeanne, U.S.A.
 Wrona, Włodzimierz S., U.S.A.
 Wu, Fang, People's Republic
 of China
 Wu, Ling-Erl Eileen T., U.S.A.
 Wu, Pei Yuan, China-Taiwan
 Wu, Wen-Tsun, People's
 Republic of China
 Wulbert, Daniel E., U.S.A.
 Wulfsohn, Aubrey, England
 Wüstholtz, Gisbert, Federal
 Republic of Germany
 Wyler, Thomas R., Switzerland
 Xambó-Descamps, Sebastian,
 Spain
 Xia, Daoxing, U.S.A.
 Xiao Gang, Xiao, People's
 Republic of China
 Yagita, Nobuaki, Japan
 Yamanoshita, Tsuneyo, Japan
 Yamanouchi, Takehiko, U.S.A.
 Yamaoka, Kenya, U.S.A.
 Yamashita, Shinji, Japan
 Yang, Chung-Chun, U.S.A.
 Yang, DaGang, U.S.A.
 Yang, Kung-Wei, U.S.A.
 Yang, Jae-Hyun, Republic of
 Korea
 Yang, Lo, People's Republic
 of China
 Yang, Paul, U.S.A.
 Yang, Wei-Shih, U.S.A.
 Yanik, Joe, U.S.A.
 Yano, Koichi, Japan
 Yanosko, Barbara J., U.S.A.
 Yanowitch, Michael, U.S.A.
 Yaqub, Adil, U.S.A.
 Yassaee Maibodi, Hedayat, Iran
 Yasuda, Yutaka, Japan
 Yasuo, Minato, Japan
 Yates, Samuel, U.S.A.
 Yau, Chi-Ming, U.S.A.
 Yau, Shing-Tung, U.S.A.
 Yau, Stephen S.-T., U.S.A.
 Ye, Yangbo, U.S.A.
 Yen, David H. Y., U.S.A.
 Yeung, Grace Chi-Dak N.,
 U.S.A.
 Ylinen, Kari E., Finland
 Yocum, Kenneth L., U.S.A.
 Yoshihiro, Mizoguchi, Japan
 Yoshimoto, Takeshi, Japan
 Yoshizawa, Taro, Japan
 Young, Nicholas J., Scotland
 Young, Robin C., U.S.A.
 Younger, Daniel H., Canada
 Younis, Rahman, Kuwait
 Yu Wenci, People's
 Republic of China
 Yuan, Chuan Kuan, U.S.A.
 Yucas, Joseph L., U.S.A.
 Yukich, Joseph E., U.S.A.
 Zabeeh, J. G., U.S.A.
 Zadeh, Lotfi A., U.S.A.
 Zafrany, Samy, U.S.A.
 Zagier, Don B., Federal
 Republic of Germany
 Zaretti, Anna, Italy
 Zayed, Ahmed I., U.S.A.
 Zeeman, E. Christopher,
 England
 Zeeman, Mary Lou, U.S.A.
 Zehnder, Eduard J., Federal
 Republic of Germany
 Zeilberger, Doron, U.S.A.
 Zelazko, Wiesław T., Poland
 Zelditch, Steven M., U.S.A.
 Zhang Gong Qing, People's
 Republic of China
 Zhang, Xue-zao, U.S.A.
 Zhang, Zhi-Fen, People's
 Republic of China
 Zharn, P. Joseph, U.S.A.
 Zhou Xue Guang, People's
 Republic of China
 Zhou, Zhengfang, U.S.A.
 Zhou, Zhiming, People's
 Republic of China
 Zieba, Jacek J., U.S.A.
 Ziegler, Guenter M., U.S.A.
 Zielezny, Zbigniew H., U.S.A.
 Zierau, Roger, U.S.A.
 Zimmer, Horst G., Federal
 Republic of Germany
 Zimmer, Robert J., U.S.A.
 Zimmerman, Benno,
 Switzerland
 Zimmermann, Irene, Federal
 Republic of Germany
 Zinde-Walsh, Victoria M., Canada
 Zitarelli, David E., U.S.A.
 Zizza, Frank, U.S.A.
 Zokayi, Alireza, Iran
 Zorn, Paul M., U.S.A.
 Zotov, Natalia V., U.S.A.
 Zou, Hung Bin, U.S.A.
 Zubelli, Jorge P., U.S.A.
 Zucker, Steven M., U.S.A.
 Zuckerman, Gregg J., U.S.A.
 Zukor, Karen S., U.S.A.
 Zumino, Bruno, U.S.A.
 Zweifel, Paul F., U.S.A.

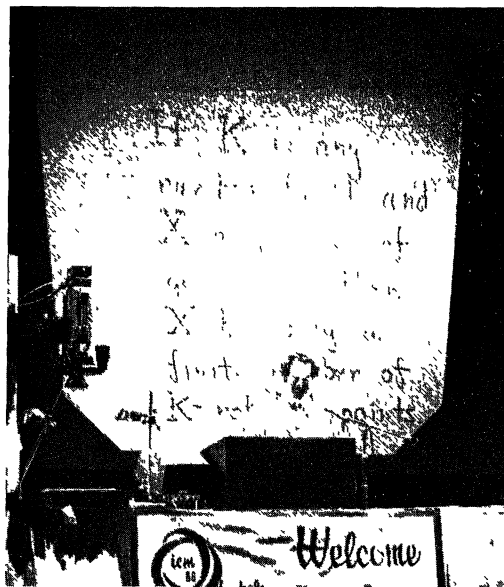
MEMBERSHIP BY COUNTRY

Algeria	3	Japan	176
Argentina	3	Jordan	1
Australia	20	Kenya	2
Austria	8	Kuwait	9
Bahrain	2	Lebanon	1
Bangladesh	1	Luxembourg	1
Belgium	20	Malaysia	3
Botswana	1	Mexico	10
Brazil	20	Netherlands	27
Bulgaria	4	New Zealand	10
Cameroon	3	Nicaragua	2
Canada	167	Nigeria	8
Chile	3	Northern Ireland	2
China-Taiwan	17	Norway	16
Colombia	2	Oman	1
Costa Rica	1	People's Republic of China	30
Cuba	3	Philippines	2
Czechoslovakia	6	Poland	17
Denmark	14	Portugal	7
Egypt	2	Romania	3
England	94	Saudi Arabia	6
Ethiopia	1	Scotland	12
Federal Republic of Germany	119	Senegal	2
Finland	19	Singapore	5
France	117	South Africa	11
German Democratic Republic	4	South Korea	6
Ghana	1	Spain	18
Greece	6	Sweden	29
Guatemala	2	Switzerland	33
Hong Kong	16	Thailand	1
Hungary	12	Trinidad	1
Iceland	6	Turkey	4
India	21	U.S.A.	2324
Indonesia	1	U.S.S.R.	57
Iran	31	U.S.V.I.	1
Iraq	4	Uruguay	2
Ireland	4	Venezuela	3
Israel	34	Vietnam	3
Italy	47	Wales	5
Ivory Coast	9	West Indies	1
Jamaica	1	Yugoslavia	10

Total 3711



Milnor speaking on Freedman's work



Mazur speaking on Faltings' work

The Work of the Medalists

The works of the Fields Medalists and the Nevanlinna Prize Winner
were presented as follows:

Fields Medalists

On the Work of *Simon K. Donaldson* by Michael Atiyah

On Some of the Mathematical Contributions of *Gerd Faltings* by B. Mazur

The Work of *Michael H. Freedman* by John Milnor

The Nevanlinna Prize Winner

The Work of *Leslie G. Valiant* by V. Strassen



Valiant, Freedman, Faltings, Donaldson

On the Work of Simon Donaldson

MICHAEL ATIYAH

In 1982, when he was a second-year graduate student, Simon Donaldson proved a result [1] that stunned the mathematical world. Together with the important work of Michael Freedman (described by John Milnor), Donaldson's result implied that there are "exotic" 4-spaces, i.e., 4-dimensional differentiable manifolds which are topologically but not differentiably equivalent to the standard Euclidean 4-space R^4 . What makes this result so surprising is that $n = 4$ is the only value for which such exotic n -spaces exist. These exotic 4-spaces have the remarkable property that (unlike R^4) they contain compact sets which cannot be contained inside any differentiably embedded 3-sphere!

To put this into historical perspective, let me remind you that in 1958 Milnor discovered exotic 7-spheres, and that in the 1960s the structure of differentiable manifolds of dimension ≥ 5 was actively developed by Milnor, Smale (both Fields Medalists), and others, to give a very satisfactory theory. Dimension 2 (Riemann surfaces) was classical, so this left dimensions 3 and 4 to be explored. At the last Congress, in Warsaw, Thurston received a Fields Medal for his remarkable results on 3-manifolds, and now at this Congress we reach 4-manifolds. I should emphasize that the stories in dimensions 3, 4, and $n \geq 5$ are totally different, with the low-dimensional cases being much more subtle and intricate.

Although I have highlighted the exotic 4-space as a spectacular corollary of the Freedman/Donaldson results, this is a by-product; their work is actually devoted to studying *closed* 4-manifolds. To such a 4-manifold, one associates standard topological invariants. In particular, for an *oriented* manifold, one gets a symmetric integer matrix of determinant ± 1 defined by the intersection properties of the 2-cycles (and depending on a choice of basis). Freedman showed that all such matrices can occur for topological 4-manifolds. Donaldson's result was that, among positive definite matrices, only those equivalent to the unit matrix can occur for differentiable 4-manifolds.¹ This is a severe restriction and shows that the differentiable and topological situations are totally different.

¹Actually, in [1] Donaldson restricted himself to simply-connected manifolds, but more recently he has succeeded in removing this restriction.

The surprise produced by Donaldson's result was accentuated by the fact that his methods were completely new and were borrowed from theoretical physics, in the form of the Yang-Mills equations. These equations are essentially a nonlinear generalization of Maxwell's equations for electro-magnetism, and they are the variational equations associated with a natural geometric functional. Differential geometers study connections and curvature in fibre bundles, and the Yang-Mills functional is just the L^2 -norm of the curvature. If the group of the fibre bundle is the circle, we get back the linear Maxwell theory, but for non-abelian Lie groups, we get a nonlinear theory. Donaldson uses only the simplest nonabelian group, namely $SU(2)$, although in principle other groups can and will perhaps be used.

Physicists are interested in these equations over Minkowski space-time, where they are hyperbolic, and also over Euclidean 4-space, where they are elliptic. In the Euclidean case, solutions giving the absolute minimum (for given boundary conditions at ∞) are of special interest, and they are called *instantons*.

Several mathematicians (including myself) worked on instantons and felt very pleased that they were able to assist physics in this way. Donaldson, on the other hand, conceived the daring idea of reversing this process and of using instantons on a general 4-manifold as a new geometrical tool. In this he has been brilliantly successful: he has unearthed totally new phenomena and simultaneously demonstrated that the Yang-Mills equations are beautifully adapted to studying and exploring this whole new field.

Of course, the use of differential equations in geometry is not new; the study of geodesics or minimal surfaces are classical examples. However, in these cases a solution of the differential equation (e.g., a minimal surface) is used as a geometrical object. Donaldson's use of instantons is quite different. I should explain that instantons as solutions of a minimization problem are not unique but typically depend on a finite number of continuous parameters, and it is the nonlinear space of these instanton parameters that Donaldson uses as a geometrical tool. The closest prior example of such an approach is the (linear) Hodge theory of harmonic forms. In fact, Hodge was directly motivated by Maxwell's equations, and instantons are a natural nonlinear generalization of harmonic forms. In the linear case the parameter space is of course linear and determined by its dimension, but in the nonlinear case there is much more information embodied in the parameter space, which is a topologically interesting manifold.

The success of Donaldson's program depends on having a thorough understanding of the analysis of the Yang-Mills equations. One needs existence, regularity, and convergence theorems, all of which are quite delicate, involving both local and global aspects. Fortunately, C. H. Taubes [6, 7] and K. Uhlenbeck [8, 9] have provided these analytical foundations, and so one can proceed to use instantons as an effective geometric tool. However, instantons cannot be bought off the shelf: to use them one has to understand and become involved with the full details of the analysis, and Donaldson has had to do this in order to put them to geometric use.

The Yang-Mills equations depend on fixing a background metric on the 4-manifold and, as in Hodge theory, Donaldson has to study the effect of varying the metric in order to derive results which depend only on the underlying manifold. Because of the nonlinearity, this is a more serious problem than in Hodge theory and great care is needed.

In fact, the Yang-Mills equations depend only on the conformal class of the metric and this conformal invariance is fundamental in physics where it implies the absence of a basic length scale. Analytically it is a source of difficulty making the equations a delicate border-line case where certain compactness arguments just fail, so that a sequence of instanton solutions can pick up Dirac delta functions in the limit. It is, however, just this delicate failure that Donaldson exploits geometrically: instead of the delta functions being regarded as undesirable singularities, they provide the key link between the 4-manifold and the instanton parameter space. One might say that the physicist's ambivalence to particles and fields is the essence of Donaldson's theory.

When Donaldson proved his first result it was by no means clear if this was some isolated case or whether instantons could be used more generally. Since then, however, Donaldson has, with great insight and skill, developed and exploited instantons with remarkable success. He has extended his results to the case of indefinite intersection matrices, providing further constraints on the topology of differentiable 4-manifolds. He has also, in the other direction, produced new invariants of 4-manifolds which can be used to distinguish smooth manifolds which are topologically equivalent. In particular, he has shown that complex algebraic surfaces (of complex dimension 2 and so of real dimension 4) appear to play a key role. In a very elegant paper [2] he proved an existence theorem which showed that, on an algebraic surface, instantons (or rather their parameter spaces) have a purely algebraic description, coinciding with what algebraic geometers call stable vector bundles. His new invariants can then be calculated algebraically and he used this [3] to exhibit two algebraic surfaces which are homeomorphic but not diffeomorphic. One of these surfaces is rational and his results strongly suggest that the rationality of an algebraic surface may be a differentiable property (it is *not* topological).

I indicated earlier that mathematicians had been working on the original physicists' problem of explicitly finding all instantons on Euclidean 4-space. In a short but decisive paper [4] Donaldson linked this problem with algebraic vector bundles on the complex projective plane (viewed as a compactification of $R^4 = C^2$). He also applied similar ideas [5] to solve a related but more difficult physical problem, that of magnetic monopoles. He proved the remarkably simple result that the parameter space of monopoles of magnetic charge k can be identified with the space of rational functions of a complex variable of degree k .

When Donaldson produced his first few results on 4-manifolds, the ideas were so new and foreign to geometers and topologists that they merely gazed in bewildered admiration. Slowly the message has gotten across and now Donaldson's ideas are beginning to be used by others in a variety of ways.

From what I have said you can see that Donaldson has opened up an entirely new area; unexpected and mysterious phenomena about the geometry of 4 dimensions have been discovered. Moreover, the methods are new and extremely subtle, using difficult nonlinear partial differential equations. On the other hand, this theory is firmly in the mainstream of mathematics, having intimate links with the past, incorporating ideas from theoretical physics, and tying in beautifully with algebraic geometry. It is remarkable and encouraging that such a young mathematician can understand and harness such a wide range of ideas and techniques in so short a time and put them to such brilliant use. It is an indication that mathematics has not lost its unity, or its vitality.

REFERENCES

1. S. K. Donaldson, *Self-dual connections and the topology of smooth 4-manifolds*, Bull. Amer. Math. Soc. **8** (1983), 81–83.
2. —, *Anti-self-dual connections over complex algebraic surfaces and stable vector bundles*, Proc. London Math. Soc. **50** (1985), 1–26.
3. —, *La topologie différentielle des surfaces complexes*, C. R. Acad. Sci. Paris **301** (1985), 317–320.
4. —, *Instantons and geometric invariant theory*, Comm. Math. Phys. **93** (1984), 453–460.
5. —, *Nahm's equations and the classification of monopoles*, Comm. Math. Phys. **96** (1984), 387–407.
6. C. H. Taubes, *Self-dual connections on non-self-dual 4-manifolds*, J. Differential Geom. **17** (1982), 139–170.
7. —, *Self-dual connections on manifolds with indefinite intersection matrix*, J. Differential Geom. **19** (1984), 517–560.
8. K. K. Uhlenbeck, *Connections with L^p bounds on curvature*, Comm. Math. Phys. **83** (1982), 11–30.
9. —, *Removable singularities in Yang-Mills fields*, Comm. Math. Phys. **83** (1982), 31–42.

UNIVERSITY OF OXFORD, MATHEMATICAL INSTITUTE, OXFORD OX1 3LB, ENGLAND

On Some of the Mathematical Contributions of Gerd Faltings

B. MAZUR

One of the recent great moments in mathematics was when Gerd Faltings revealed the circle of ideas which led him to a proof of the conjecture of Mordell ([1]; see also [2, 3]).

The conjecture, marvelous in the simplicity of its statement, had stood as a goad and an elusive temptation for over half a century: it is even older than the Fields Medal! In modern language it takes the following form:

If K is any number field and X is any curve of genus > 1 defined over K , then X has only a finite number of K -rational points.

To get a feeling for our level of ignorance in the face of such questions, consider that, before Faltings, there was not a single curve X (of genus > 1) for which we knew this statement to be true for all number fields K over which X is defined!

Already in the twenties, Weil and Siegel made serious attempts to attack the problem. Siegel, influenced by Weil's thesis, used methods of diophantine approximation, to prove that the number of *integral* solutions to a polynomial equation $f(X, Y) = 0$ (i.e., solutions in the ring of integers of a number field K) is finite, provided that f defines a curve over K of genus > 0 , or a curve of genus 0 with at least three points at infinity.

In his thesis, Weil generalized Mordell's theorem on the finite generation of the group of rational points on an elliptic curve, to abelian varieties of any dimension. Weil then hoped to use this finite generation result for the rational points on the jacobian of a curve to go on to show that when a curve of genus > 1 is imbedded in its jacobian, only a finite number of the rational points of the jacobian can lie on the curve. Not finding a way to do this, he decided to call his proof of finite generation (the "theorem of Mordell-Weil") a thesis, despite Hadamard's advice not to be satisfied with half a result!

After this work of Weil and Siegel there was little progress for thirty years. It was in the sixties and early seventies that several new developments occurred in algebraic geometry and number theory which were to influence Faltings (work of

Grothendieck, Serre, Mumford, Lang, Néron, Tate, Manin, Shafarevich, Parsin, Arakelov, Zarhin, Raynaud, and others).

These developments, which enter in an essential way in the work of Faltings, encompass three grand mathematical themes, and Faltings proved the conjecture of Mordell, by first establishing the truth of some other outstanding conjectures—fundamental to arithmetic and to arithmetic algebraic geometry. In the next few minutes I should like to try to convey some sense of the sweep of Faltings's accomplishment by touching on those themes, and those conjectures.

1. The arithmetization of geometry and the geometrization of arithmetic. Nowadays the analogy between number fields and fields of rational functions on algebraic curves (over finite fields) is so well imprinted upon our view of both number theory and the theory of algebraic curves that it is hard to imagine how we might deal with either theory, if deprived of that analogy. A firm understanding of the power of such analogies was present already in the work of Kronecker, and of Dedekind and Weber at the end of the last century. This understanding was deepened by the development of algebraic number theory, in the hands of Artin and Chevalley, of algebraic geometry, via the foundations developed by Zariski, Weil, and Serre, and more recently, of arithmetic algebraic geometry, whose foundations are given by the language of schemes, of Grothendieck.

In the language of schemes, a smooth curve over a finite field and the ring of integers in a number field are not merely analogous: they are two instances of the same notion (*regular schemes of dimension one, of finite type over \mathbb{Z}*). Similarly, a family of curves over a “base” curve over a finite field and a curve over the ring of integers in a number field are companion instances of the same notion.

This is not to say that this analogy is thoroughly understood! Why, it has only been relatively recently, thanks to the groundbreaking work of Arakelov, that we have begun to see a format for bringing the *archimedean places* of a number field into the geometric picture.

Moreover, this “synthetic view,” extraordinarily efficacious for carrying problems and conjectures from the realm of function fields to the realm of number fields and back again, is far less satisfactory when it comes to carrying the *proofs* of those conjectures from one realm to the other.

For example, the analogue of Mordell's conjecture in the function field case was first settled by Manin back in 1963. A different proof was given by Grauert in 1965. Arakelov (using a beautiful idea of Parsin) found another proof in 1971. Yet another proof, also using Parsin, was given by Zarhin in 1974.

Even when we were armed with these approaches to the function field case, the number field case seemed intractable for almost a decade until Faltings discovered a method, analogous to that of Parsin-Zarhin, which established the classical conjecture of Mordell over number fields. To this day we lack number field analogues of the other approaches to the problem—say, of Manin's

original proof, or of Arakelov's (Do they exist?). Judging from Faltings's published contributions to Arakelov's theory [4] one might imagine that he himself had been simultaneously pursuing two approaches to the Mordell conjecture in the number field case: one suggested by Arakelov's work, and the other by Zarhin's. The problem yielded, in 1983, to the second approach.

Faltings's method in the number field case, and Zarhin's in the function field case, was to settle Mordell's conjecture by first answering a more "geometric" question:

Kodaira had raised the problem of studying, or perhaps "classifying," all (truly varying) families of smooth curves of a given genus over a fixed (not necessarily complete) base curve. It was Shafarevich, at the 1962 International Congress, who first brought attention to the analogue of Kodaira's problem in arithmetic, and to its significance. One version of this analogue, now known as *Shafarevich's Conjecture for Curves*, states that

There are only a finite number of nonisomorphic curves of a given genus > 1 defined over a fixed number field and possessing good reduction outside a fixed finite set of primes in the ring of integers of that number field.

One way of paraphrasing Kodaira's original problem is by formulating the "function field analogue" of Shafarevich's conjecture, where the ring of integers in a number field is replaced by a base curve over a finite field. By an ingenious argument which happily worked as well in the number field case as in the function field case, Parsin had shown in 1968 that Mordell's conjecture follows from Shafarevich's conjecture.

Very roughly, Parsin's idea is as follows: Fix $g > 1$. Given a curve X of genus g and a rational point P on X over a number field K , Parsin produced a converging Y of X which is ramified only above P , and whose number field of definition and set of bad primes are "uniformly bounded" in terms of the data: g , the number field of definition of X and P , and the set of bad primes of X . Since Y determines the pair (X, P) up to finite ambiguity, it follows that if there are only finitely many such Y 's (*Shafarevich's conjecture*) then there only finitely many such P 's (*Mordell's conjecture*).

In their respective contexts, both Zarhin's and Faltings's attack on Mordell's conjecture is to prove Shafarevich's conjecture for curves (over function fields, and over number fields, respectively), and then to appeal to Parsin's idea.

2. Curves and abelian varieties. It was Weil, in his proof of the "Riemann hypothesis for curves over finite fields," who first made essential use of the passage from curves to abelian varieties to derive important consequences for the arithmetic of curves.

The "geometric" insight, that in pursuing questions about curves it sometimes pays to appeal to their jacobians, goes further back. Indeed, the fact that we see such a close relationship between curves and their jacobians is one of our many legacies from the Italian school of algebraic geometry.

With the trend towards the arithmetization of geometry, it was natural to study closely models for the jacobians of curves, or more generally for abelian varieties, over the rings of integers of number fields. Néron, in 1964, discovered the remarkable, and remarkably useful, fact that any abelian variety over a number field has a “best” model over the ring of integers of the number field—“best,” from the point of view of niceness of the reduction of the model modulo prime ideals in the ring. Although Néron, at the time he did his work, was not aware of Kodaira’s contributions in the complex analytic case, one may view Néron’s theory as a far-reaching amplification and arithmetization of a program initiated by Kodaira.

Néron models now play an important role in any close arithmetic study of abelian varieties, and in particular, they play a role in the detailed analysis of compactifications of moduli spaces for abelian varieties. The systematic *arithmetic* study of moduli spaces and their compactifications—a study initiated by the magnificent work of Mumford—in turn plays a key role in Faltings’s approach. Compactification of moduli spaces of abelian varieties over \mathbf{Z} , incidentally, is a subject to which Faltings has returned more recently: By refining, in certain respects, the work of Ching-Li Chai, Faltings has clarified some questions of arithmetic compactifications, and has thereby significantly simplified, and rendered more natural, the logical structure of his proof of Mordell’s conjecture. In his initial proof [1] Faltings took a somewhat more circuitous route, using moduli spaces of curves rather than of abelian varieties (and the published account of the technical issues was supremely succinct, requiring a certain expertise on the part of the reader).

Thanks to the close relationship between curves and their jacobians (the classical theorem of Torelli plus a finiteness result concerning polarizations), Shafarevich’s conjecture for curves (of genus > 1) reduces to a similar conjecture (also called *Shafarevich’s conjecture*) for abelian varieties:

There are only a finite number of abelian varieties of fixed dimension over a fixed number field whose Néron models possess good reduction outside a fixed finite set of primes of the number field.

This conjecture was also settled affirmatively by the work of Faltings. It was settled in tandem with another basic arithmetic question:

3. Abelian varieties and Galois representations. It was by considering number-theoretic analogies of the classical conjectures of Hodge (concerning algebraic cycles) and geometric analogues of the conjecture of Birch and Swinnerton-Dyer that Tate in 1963 formulated the following conjecture, which links the problem of classifying abelian varieties (up to isogeny) to that of classifying the Galois representations to which they give rise:

Let l be a prime number. Let K be a number field, and \overline{K} an algebraic closure of K . An abelian variety A over K is determined up to isogeny (over K) by the

natural representation of $\text{Gal}(\overline{K}/K)$ on the Q_l -vector space

$$V_l(A) = \text{Hom}(Q_l, A_l(K)),$$

where $A_l(K)$ denotes the group of \overline{K} -valued points of A , of l -power order.

This clearly basic ‘link’ between abelian varieties (denizens of algebraic geometry) and Galois representations (ostensibly more ‘elementary’ creatures) was proved by Tate himself in 1967 for a finite field K , but over number fields it was not even known to be the case for elliptic curves, before the work of Faltings.

By a beautiful argument (which makes crucial use of two theories upon which we shall comment in a moment, and) which shuttles back and forth between the conjecture of Shafarevich for abelian varieties and the conjecture of Tate, Faltings showed that both of these conjectures are true. The phrase “shuttles back and forth” is quite inadequate to characterize Faltings’s mode of argument, which captures in its weave all the mathematical themes upon which I have touched.

The “two theories” referred to are the *Theory of Heights* and the *Theory of p -Divisible Groups* (and, more generally, of *Group Schemes of Exponent p*).

The *Theory of Heights* was initiated by Weil in 1928 as a technique for “counting” rational points on abelian varieties and was used in an essential manner in his proof of the theorem “of Mordell-Weil.” This theory was further developed by Néron, by Tate, and more recently was given a new twist in the work of Arakelov.

The *Theory of p -Divisible Groups* was invented by Serre and Tate, and, independently, by Barsotti in the mid-sixties to provide a technique to analyze the way in which p -power torsion points on abelian varieties “degenerate” when specialized to characteristic p . In 1966 Tate proved the analogue of his conjecture on abelian varieties for p -divisible groups over a local field of characteristic 0. Faltings uses this theorem, and, moreover, makes essential use of an important refinement of it (covering the case of group schemes of exponent p) due to Raynaud.

Even the above recitation does not completely exhaust the list of longstanding conjectures established by Faltings in the course of his work on the Mordell conjecture. For example, an important adjunct to the conjecture of Tate concerning the representations of the Galois group $\text{Gal}(\overline{K}/K)$ on l -power torsion points of an abelian variety A defined over a number field K is the assertion of *semisimplicity* of the representation of $\text{Gal}(\overline{K}/K)$ on $V_l(A)$. This *Semisimplicity Conjecture* has also been proved by Faltings (as its “function field analogue” had been proved by Zarhin in the course of his work). Moreover the proof of semisimplicity plays a structural role in the proof of the other conjectures.

The above *Semisimplicity Conjecture* had been formulated by Grothendieck as the “dimension one case” of a more general question (semisimplicity of Galois representations acting on the d -dimensional l -adic cohomology of irreducible smooth projective varieties over number fields, for any d). One could find support for Grothendieck’s conjecture, at the time he made it, in the work

of Serre concerning the richness of the action of Galois on the torsion points of elliptic curves defined over number fields. Thanks to Faltings, we now know the *Semisimplicity Conjecture of Grothendieck* to be true for $d = 1$. This result of Faltings, incidentally, together with a technique coming from Faltings's proof, has very recently been used by Serre in a deep study of the action of Galois on torsion points of abelian varieties of arbitrary dimension g , defined over number fields.

The general case of Grothendieck's conjecture (i.e., for $d > 1$) is still open.

We have been discussing only Gerd Faltings's approach to the conjecture of Mordell, but his other mathematical contributions, whether they be concerned with moduli spaces of abelian varieties, the Riemann-Roch theorem for arithmetic surfaces, or p -adic Hodge theory, all immediately impress one as the work of a marvelously original mind from which we may expect similarly wonderful things in the future.

REFERENCES

1. G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), 349–366.
2. G. Faltings, G. Wustholtz, et al., *Rational points* (Seminar Bonn/Wuppertal, 1983/84), Aspects of Math., vol. E6, Vieweg, Braunschweig-Wiesbaden, 1984.
3. L. Szpiro, *Séminaire sur les pinceaux arithmétiques: La conjecture de Mordell*, Astérisque, no. 127, Soc. Math. France, Paris, 1985.
4. G. Faltings, *Calculus on arithmetic surfaces*, Ann. of Math. (2) **119** (1984), 387–424.

HARVARD UNIVERSITY, CAMBRIDGE, MASSACHUSETTS 02138, USA

The Work of M. H. Freedman

JOHN MILNOR

Michael Freedman has not only proved the Poincaré hypothesis for 4-dimensional topological manifolds, thus characterizing the sphere S^4 , but has also given us classification theorems, easy to state and to use but difficult to prove, for much more general 4-manifolds. The simple nature of his results in the topological case must be contrasted with the extreme complications which are now known to occur in the study of differentiable and piecewise linear 4-manifolds.

The “ n -dimensional Poincaré hypothesis” is the conjecture that every topological n -manifold which has the same homology and the same fundamental group as an n -dimensional sphere must actually be homeomorphic to the n -dimensional sphere. The cases $n = 1, 2$ were known in the nineteenth century, while the cases $n \geq 5$ were proved by Smale, and independently by Stallings and Zeeman and by Wallace, in 1960–61. (The original proofs needed an extra hypothesis of differentiability or piecewise linearity, which was removed by Newman a few years later.) The 3- and 4-dimensional cases are much more difficult.

Freedman’s 1982 proof of the 4-dimensional Poincaré hypothesis was an extraordinary tour de force. His methods were so sharp as to actually provide a complete classification of all compact simply connected topological 4-manifolds, yielding many previously unknown examples of such manifolds, and many previously unknown homeomorphisms between known manifolds. He showed that a compact simply connected 4-manifold M is characterized, up to homeomorphism, by two simple invariants. The first is the 2-dimensional homology group

$$H_2 = H_2(M; Z) \cong Z \oplus \cdots \oplus Z,$$

together with the symmetric bilinear intersection pairing

$$\omega: H_2 \otimes H_2 \rightarrow Z.$$

This pairing, which is defined as soon as we choose an orientation for M , must have determinant ± 1 by Poincaré duality. The second is the Kirby-Siebenmann obstruction class, an element

$$\sigma \in H^4(M; Z/2) \cong Z/2$$

that vanishes if and only if M is stably smoothable. In other words, σ is zero if and only if the product $M \times R$ can be given a differentiable structure, or

equivalently a piecewise linear structure. These two invariants ω and σ can be prescribed arbitrarily, except for a relation in one special case. If the form ω happens to be even, that is if $\omega(x, x) \equiv 0 \pmod{2}$ for every $x \in H_2$, then the Kirby-Siebenmann obstruction must be equal to the Rohlin invariant:

$$\sigma \equiv \text{signature}(\omega)/8 \pmod{2}.$$

(Freedman's original proof that these two invariants characterize M up to homeomorphism required an extra hypothesis of "almost-differentiability," which was later removed by Quinn.)

If the intersection form $\omega \neq 0$ is indefinite or has rank at most eleven, then it follows from known results about quadratic forms that M can be built up (nonuniquely) as a connected sum of copies of four simple building blocks, each of which may be given either the standard or the reversed orientation. One needs the product $S^2 \times S^2$, the complex projective plane CP^2 , and two exotic manifolds which were first constructed by Freedman. One of these is a nondifferentiable analogue of the complex projective plane, and the other is the unique manifold whose intersection form ω is positive definite and even of rank eight. (This ω can be identified with the lattice generated by the root vectors of the Lie group E_8 . As noted by Rohlin in 1952, a 4-manifold with such an intersection form can never be differentiable.) By way of contrast, if we allow positive definite intersection forms, then the number of distinct simply connected manifolds grows more than exponentially with increasing middle Betti number.

Freedman's methods extend also to noncompact 4-manifolds. For example, he showed that the product $S^3 \times R$ can be given an exotic differentiable structure, which contains a smoothly embedded Poincaré homology 3-sphere and hence cannot be smoothly embedded in euclidean 4-space [11, 16]. His methods apply also to many manifolds which are not simply connected [22]. For example, a "flat" 2-sphere in 4-space is unknotted if and only if its complement has free cyclic fundamental group; and a flat 1-sphere in S^3 has trivial Alexander polynomial if and only if it bounds a flat 2-disk in the unit 4-disk whose complement has free cyclic fundamental group.

The proofs of these results are extremely difficult. The basic idea, which had been used in low dimensions by Moebius and Poincaré, and in high dimensions by Smale and Wallace, is to build the given 4-manifold up inductively, starting with a 4-dimensional disk, by successively adding handles. The essential difficulty, which does not arise in higher dimensions, occurs when we try to control the fundamental group by inserting 2-dimensional handles, since a 2-dimensional disk immersed in a 4-manifold will usually have self-intersections. This problem was first attacked by Casson, who showed how to construct a generalized kind of 2-handle with prescribed boundary within a given 4-manifold. Freedman's major technical tool is a theorem which asserts that every Casson handle is actually homeomorphic to the standard open handle, (closed 2-disk) \times (open 2-disk). The proof involves a delicately controlled infinite repetition argument in the spirit of the Bing school of topology, and is nondifferentiable in an essential way.

PAPERS OF M. H. FREEDMAN

1. *Automorphisms of circle bundles over surfaces*, Geometric Topology, Lecture Notes in Math., vol. 438, Springer-Verlag, 1974, pp. 212–214.
2. *On the classification of taut submanifolds*, Bull. Amer. Math. Soc. **81** (1975), 1067–1068.
3. *Uniqueness theorems for taut submanifolds*, Pacific J. Math. **62** (1976), 379–387.
4. *Surgery on codimension 2 submanifolds*, Mem. Amer. Math. Soc. No. 119 (1977).
5. *Une obstruction élémentaire à l'existence d'une action continue de groupe dans une variété* (with W. Meeks), C. R. Acad. Sci. Paris Ser. A **286** (1978), 195–198.
6. *Λ -splitting 4-manifolds* (with L. Taylor), Topology **16** (1977), 181–184.
7. *A geometric proof of Rochlin's theorem* (with R. Kirby), Proc. Sympos. Pure Math., vol. 32, part 2, Amer. Math. Soc., Providence, R.I., 1978, pp. 85–98.
8. *Remarks on the solution of first degree equations in groups*, Algebraic and Geometric Topology, Lecture Notes in Math., vol. 664, Springer-Verlag, 1978, pp. 87–93.
9. *Quadruple points of 3-manifolds in S^4* , Comment. Math. Helv. **53** (1978), 385–394.
10. *A converse to (Milnor-Kervaire theorem) $\times R$ etc...*, Pacific J. Math. **82** (1979), 357–369.
11. *A fake $S^3 \times R$* , Ann. of Math. **110** (1979), 177–201.
12. *Cancelling 1-handles and some topological imbeddings*, Pacific J. Math. **80** (1979), 127–130.
13. *A quick proof of stable surgery* (with F. Quinn), Comment. Math. Helv. **55** (1980), 668–671.
14. *Planes triply tangent to curves with nonvanishing torsion*, Topology **19** (1980), 1–8.
15. *Slightly singular 4-manifolds* (with F. Quinn), Topology **20** (1981), 161–173.
16. *The topology of 4-manifolds*, J. Differential Geom. **17** (1982), 357–454 (see also F. Quinn, *ibid.*, p. 503).
17. *A surgery sequence in dimension four; the relations with knot concordance*, Invent. Math. **68** (1982), 195–226.
18. *Closed geodesics on surfaces* (with J. Hass, P. Scott), Bull. London Math. Soc. **14** (1982), 385–391.
19. *A conservative Dehn's Lemma*, Low Dimensional Topology, Contemp. Math., vol. 20, Amer. Math. Soc., Providence, R.I., 1983, pp. 121–130.
20. *Imbedding least area incompressible surfaces* (with J. Hass, P. Scott), Invent. Math. **71** (1983), 609–642.
21. *Homotopically trivial symmetries of Haken manifolds are toral* (with S.-T. Yau), Topology **22** (1983), 179–189.
22. *The disk theorem for 4-dimensional manifolds*, Proc. Internat. Congr. Math. Warsaw 1983, vol. 1, Pol. Sci. Publ., 1984, pp. 647–663.
23. *There is no room to spare in four-dimensional space*, Notices Amer. Math. Soc. **31** (1984), 3–6.
24. *Atomic surgery problems* (with A. Casson), Four-Manifold Theory, Contemp. Math., vol. 35, Amer. Math. Soc., Providence, R.I., 1984, pp. 181–199.

OTHER REFERENCES

25. H. Poincaré, *Cinquième complément à l'Analysis situs* (1904), Oeuvres 6, Paris, 1953, pp. 435–498.
26. L. Siebenmann, *Amorces de la chirurgie en dimension 4, un $S^3 \times R$ exotique* (d'après A. Casson et M. H. Freedman), Séminaire Bourbaki (1978–79), Exp. No. 536, Lecture Notes in Math., vol. 770, Springer-Verlag, 1980, pp. 183–207.
27. —, *La conjecture de Poincaré topologique en dimension 4* (d'après M. H. Freedman), Séminaire Bourbaki (1981–82), Exp. No. 588, Astérisque, no. 92–93, Soc. Math. France, Paris, 1982, pp. 219–248.

The Work of Leslie G. Valiant

V. STRASSEN

Theoretical computer science is very young, when compared for instance to number theory, geometry, or topology. While these classical fields are like magnificent old oaks, whose growth takes place at dizzying heights and is therefore not easy to follow, theoretical computer science resembles a fast-growing young tree, whose fresh green may be perceived and enjoyed by everyone coming near.

Leslie G. Valiant has contributed in a decisive way to the growth of almost every branch of this young tree. In order to convey some impression of the scope of his work and the impetuous pace of its creation, I shall first discuss three early papers, published within a single year, which contain spectacular advances in three very different areas. Then I will turn to what is perhaps Valiant's most important and mature work, centering around his theory of counting problems.

Languages. *Context-free grammars* were introduced by N. Chomsky in 1956 as a means of analyzing natural languages. They are now being used extensively for describing the structure of programming languages. The central algorithmic problem is to recognize sentences of the language defined by such a grammar. For a number of years recognition algorithms which run in a time proportional to n^3 on sentences of length n had been known, but in spite of much effort no significant improvement had been obtained except in special cases.

In a coup de main Valiant showed in 1975 that the recognition problem can be solved in less than cubic time by reducing it to integer matrix multiplication. He has never come back to this subject, but ingenious algorithmic reductions from combinatorial to algebraic problems have become one of the main themes of his work.

Graphs. Let m be a positive integer. An m -*superconcentrator* is a directed graph with m input and m output nodes, such that for every $r \leq m$ any r input nodes may be connected to any r output nodes in some order by r disjoint directed paths. By the size of an m -superconcentrator, one means its number of edges.

Superconcentrators first appeared in algebraic complexity theory: Any straightline algorithm for computing the Discrete Fourier Transform of order m

yields an m -superconcentrator of a size proportional to the length of the algorithm. In particular, the Fast Fourier Transform provides examples of m -superconcentrators of size $\text{const} \cdot m \log m$ and any improvement of the Fast Fourier Transform would lead to m -superconcentrators of still smaller size. Aho, Hopcroft, and Ullman, among others, stated the problem of proving or disproving that the minimal size of m -superconcentrators grows like $m \log m$. By the previous remarks a positive answer would yield an optimality proof for the Fast Fourier Transform up to order of magnitude.

Valiant dashed such hopes by showing that there exist m -superconcentrators of size linear in m . Considering the flexibility of these graphs the result is almost unbelievable. Valiant's proof, which is based upon previous work of Pinsker, combines a counting argument with an elegant recursive construction. (Let me note that Margulis and Gabber-Galil have succeeded in replacing the counting argument by an explicit construction as well.) In the decade after their discovery superconcentrators of linear size have become useful tools in the information and communication sciences far beyond their original purpose.

This work is just one example of Valiant's systematic and penetrating study of efficient imbedding and routing properties of graphs, leading in recent years to a theory of the general purpose parallel computer (the so-called supercomputer).

Turing machines. Since their invention by A. Turing in 1936 Turing machines have been the principal theoretical model on which notions of computability and computational complexity have been based. Often a Turing machine is used as a decision procedure for some property of numbers, graphs, logical formulas, etc., which by a suitable encoding are presented to the machine as binary strings. While recursion theory is concerned with the decidability of decision problems, complexity theory also considers the amount of time and space needed to reach a decision.

Given a function $t: \mathbb{N} \rightarrow \mathbb{N}$, let $\text{TIME}(t)$ be the class of all decision problems (i.e., sets of binary strings) that can be decided by a multitape Turing machine using only $O(t(n))$ computational steps on inputs of length n . Define $\text{SPACE}(t)$ similarly in terms of the number of tape squares visited. Obviously $\text{TIME}(t) \subset \text{SPACE}(t)$, since in one step a Turing machine can reach at most a constant number of new tape squares.

The dualism of time and space, which to a large extent shapes the physical sciences, is also present in the discrete world of idealized computers. As a first but fundamental question we may ask whether the above inclusion is strict:

$$\text{TIME}(t) \subsetneq \text{SPACE}(t)? \quad (\text{A})$$

— Computing experience overwhelmingly indicates that this is indeed the case, but — there does not seem to exist an easy proof. It was a scientific sensation when Hopcroft, Paul, and Valiant showed that in fact one has

$$\text{TIME}(t) \subset \text{SPACE}(t/\log t), \quad (\text{B})$$

which implies (A), since $\text{SPACE}(t/\log t)$ is strictly contained in $\text{SPACE}(t)$ by a standard diagonalization argument. (Alluding to H. Weyl's classic the result might be stated: Space or time—it matters.)

Hopcroft, Paul, and Valiant reduced the proof of (B) to finding a good strategy for a certain game on graphs, the “pebble game,” which had been invented earlier by Paterson and Hewitt for a different purpose. This strategy was later shown to be optimal by Paul, Tarjan, and Celoni, so that no improvement of (B) may be expected by the same method. (It is a pleasant coincidence that this optimality proof uses linear sized superconcentrators as an essential ingredient.)

The result of Hopcroft, Paul, and Valiant is weak in the following sense: The complexity classes $\text{TIME}(t)$ and $\text{SPACE}(t)$ depend on the Turing machine model of computation and therefore the validity of (A) or (B) may be lost by a change of models. A more robust formulation of the problem of time and space involves the complexity classes

$$\text{PTIME} = \bigcup_k \text{TIME}(n^k) \quad \text{and} \quad \text{PSPACE} = \bigcup_k \text{SPACE}(n^k)$$

of all problems decidable in polynomial time or space. There is no doubt that the inclusion

$$\text{PTIME} \subset \text{PSPACE}$$

is also strict, but this has not yet been proven.

The class PTIME, or simply P, as it is called, has gained a central position in complexity theory, since it appears to be best suited for distinguishing between what can be and what cannot be computed in practice. For brevity I shall call the problems in P easy, those not in P hard.

Complete problems. Consider a map $f: \mathbb{N} \rightarrow 2^{\mathbb{N}}$ such that

$$\{(x, y) : y \in f(x)\} \quad \text{is easy,} \tag{1}$$

$$y \in f(x) \Rightarrow |y| \leq |x|^k \tag{2}$$

for a suitable constant k . ($|x|$ denotes the binary length of x). The set of all numbers x such that $f(x)$ is nonempty is called a *search problem*. The complement of the set of primes is an example: $f(x)$ may be taken to consist of all proper divisors of x . The name “search problem” refers to the possibility of searching through all y satisfying the inequality in (2) for an element of $f(x)$. However, although each test of membership is easy by (1), such an exhaustive search may use exponential time due to the number of tests to be conducted.

The class NP of all search problems lies between P and PSPACE. It turns out that a great number of decision problems occurring in mathematics and its applications belong to NP, when suitably encoded. Here are a few examples: to decide

- if a propositional formula is satisfiable,
- if a graph is isomorphic to a subgraph of another,
- if a graph has a Hamilton circuit,

if a graph has a 3-coloring,

if a diophantine equation of the form $F(X_1, \dots, X_n) = c$, where c and the coefficients of F are natural numbers, has a solution in natural numbers.

In his seminal paper, "The Complexity of Theorem Proving Procedures," S. Cook in 1971 showed that the satisfiability problem as well as the subgraph problem are *complete* in the class NP under polynomial reduction, in short, NP-complete. Roughly speaking, NP-complete problems have maximal degree of difficulty among all search problems. It follows that if either the satisfiability or the subgraph problem is easy, then every search problem is easy.

A more precise formulation of Cook's theorem implies an even stronger statement; namely, that a fast algorithm for deciding the satisfiability or the subgraph problem would create a fast decision procedure for any effectively given search problem in a completely mechanical way. Thus it could be used as a masterkey for search problems from all branches of mathematics. For instance, no additional competence in number theory would be needed for designing a fast test of primality or of representability of a number by any given positive polynomial. This appears so unlikely that there is little doubt about the validity of what we call *Cook's hypothesis*; namely, that the satisfiability and the subgraph problems are indeed hard, or equivalently that

$$P \neq NP.$$

The list of problems proved NP-complete was significantly extended by R. Karp in 1972 to include among others the Hamilton circuit and the graph coloring problems above. By now, most of the naturally occurring search problems have been classified as either easy or NP-complete. (The solvability of positive diophantine equations is NP-complete, even if it is restricted to binary quadratic polynomials, as has been shown by Manders and Adleman.)

Let me turn to Valiant's work on the subject. Given a map $f: \mathbb{N} \rightarrow 2^{\mathbb{N}}$ as in the definition of a search problem, we may not only ask whether $f(x)$ is nonempty, but may inquire about its size. Valiant calls the function $x \mapsto \#f(x)$ a *counting problem* and shows that the class of all counting problems also contains complete members, for instance, the counting problems corresponding to the NP-complete search problems of our list. Since counting solutions is at least as difficult as deciding whether a solution exists, complete counting problems are hard under Cook's hypothesis. Most exciting is Valiant's discovery in 1979 of various complete counting problems that correspond to easy search problems, thereby considerably enlarging the scope of the theory of NP-completeness. I give three examples:

(I) Counting subtrees of a directed graph. Note that if "subtrees" is replaced by "spanning subtrees," the counting problem becomes easy in view of a variant of a classical theorem of Kirchhoff (1847).

(II) Evaluating the probability of failure of an unreliable connecting network. According to Laplace's definition of probability as a proportion, this is tantamount to a counting problem.

(III) Counting perfect matchings of a bipartite graph. Here the corresponding search problem is not trivial as in I and II, but it is easy by a well-known algorithm of M. Hall.

In Valiant's treatment III is the critical example. The proof of completeness of this problem intricately combines ideas belonging to mathematical logic, graph theory, and algebra.

The number of perfect matchings of a bipartite graph is equal to the value of the permanent function at the zero-one-matrix representing the graph—over the ring of integers. What happens if we replace \mathbf{Z} by $\mathbf{Z}/m\mathbf{Z}$? When $m = 2$ the permanent coincides with the determinant and is therefore easy to compute. When m is any fixed power of 2, Valiant shows that the problem remains easy. In contrast, he obtains the wonderful result that the existence of a fast algorithm for the permanent modulo m for some m that is not a power of 2 implies that any polynomially bounded number-theoretical function, whose graph is easy to decide, is itself easy to compute. ("Polynomially bounded" is to be understood in terms of lengths.) It can be deduced that if the permanent modulo 3 (say) is easy, then so is prime factorization of integers.

The permanent is a polynomial function of the entries in the matrix. Thus Valiant was naturally led into *algebraic complexity* theory, which we have already touched upon when discussing superconcentrators. Here the basic model is that of a straightline algorithm, i.e., a finite sequence of arithmetical instructions, to be executed over a suitable algebraic structure. For a number of years it had seemed that this subject would remain unaffected by the notion of NP-completeness. Motivated by his work on the permanent, however, Valiant developed a convincing analogue of the theory of search and counting problems entirely in the algebraic framework. The new theory differs considerably from its model, and it will not be possible to describe even its main features here. However, the counterpart of Cook's hypothesis—let us call it Valiant's hypothesis—may be sandwiched between two succinct statements of a classical algebraic flavor: Fix a field F of characteristic $\neq 2$ and let $t(n)$ be the smallest number r such that the permanent of size n may be obtained from the determinant of size r by a simple substitution, i.e., one in which variables may be replaced only by variables or elements of F . Then Valiant's hypothesis implies that $t(n)$ grows faster than any power of n . In turn, if $t(n)$ grows even faster than $e^{(\log n)^q}$ for some q (for instance, if it grows exponentially), Valiant's hypothesis is true over the field F .

For some of you it may seem that the theories discussed here rest on weak foundations. They do not. The evidence in favor of Cook's and Valiant's hypotheses is so overwhelming, and the consequences of their failure are so grotesque, that their status may perhaps be compared to that of physical laws rather than that of ordinary mathematical conjectures. Nevertheless a traditional proof would be of great interest, and it seems to me that Valiant's hypothesis may be easier to confirm than Cook's (for example, by using the powerful methods of algebraic geometry).

The selection of Valiant's works that I have presented to you does not do justice to many of his other achievements of comparable importance: dealing with boolean complexity, probabilistic algorithms, monotone and parallel computation, artificial intelligence. They all give ample evidence of his astuteness, originality, and taste.

Theoretical computer science is in the stage of formulating its central problems and devising the proof techniques for their solution. Valiant has been eminently involved in this process, not only by answering a number of recalcitrant open questions, but above all by developing important new concepts, which have led him to discover deep and beautiful connections between problems that had seemed to be totally unrelated.

In every scientific discipline, finding fruitful concepts is a most demanding task. This is especially true for a new field, since there exist so many conceptual possibilities. But the rewards balance the difficulties: the young shoots of the sapling will become the main boughs of the full grown tree.

I have no doubt that this applies to the work of Leslie Valiant. Let me wish him, and the three Fields medalists, futures as bright as their scientific pasts.

SELECTED PUBLICATIONS OF L. G. VALIANT

1. *The equivalence problem for deterministic finite-turn pushdown automata*, Inform. and Control **25** (1974), 123–133.
2. *Parallelism in comparison problems*, SIAM J. Comput. **4** (1975), 348–355.
3. *General context-free recognition in less than cubic time*, J. Comput. System Sci. **10** (1975), 308–315.
4. *On non-linear lower bounds in computational complexity*, Seventh Annual ACM Symposium on Theory of Computing (Albuquerque, N.M., 1975), Assoc. Comput. Mach., New York, 1975, pp. 45–53. (See also J. Comput. System Sci. **13** (1976), 278–285.)
5. *On time versus space and related problems* (with J. E. Hopcroft and W. J. Paul), 16th Annual Symposium on Foundations of Computer Science (Berkeley, Calif., 1975), IEEE Computer Society, Long Beach, Calif., 1975, pp. 57–64. (See also J. Assoc. Comput. Mach. **24** (1977), 332–337.)
6. *Graph-theoretic arguments in low-level complexity*, Lecture Notes in Comput. Sci., vol. 53, Springer-Verlag, 1977, pp. 162–176.
7. *Fast probabilistic algorithms for Hamiltonian circuits and matchings* (with D. Angluin), J. Comput. System Sci. **18** (1979), 155–193.
8. *The complexity of computing the permanent*, Theoret. Comput. Sci. **8** (1979), 189–201.
9. *The complexity of enumeration and reliability problems*, SIAM J. Comput. **8** (1979), 410–421.
10. *Completeness classes in algebra*, Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing (Atlanta, Ga., 1979), Assoc. Comput. Mach., New York, 1979, pp. 249–261.
11. *Negation can be exponentially powerful*, Theoret. Comput. Sci. **12** (1980), 303–314.
12. *Reducibility by algebraic projections*, Logic and Algorithms (Zurich, 1980), Monograph. Enseign. Math., No. 30, Univ. Genève, Geneva, 1982, pp. 365–380.
13. *A scheme for fast parallel communication*, SIAM J. Comput. **11** (1982), 350–361.
14. *Universal schemes for parallel communication* (with G. J. Brebner), Proceedings of the 13th ACM Symposium on Theory of Computing (Milwaukee, Wisconsin, 1981), Assoc. Comput. Mach., New York, pp. 263–277.
15. *Fast parallel computation of polynomials using few processors* (with S. Skyum, S. Berkowitz, and C. Rackoff), SIAM J. Comput. **12** (1983), 641–644.

16. *Short monotone formula for the majority function*, J. Algorithms **5** (1984), 363–366.
17. *A theory of the learnable*, Comm. ACM **27** (1984), 1134–1142.
18. *Learning disjunctions of conjunctions*, Proceedings of Ninth International Joint Conference on Artificial Intelligence (Los Angeles, Calif., 1985), pp. 560–566.
19. *A complexity theory based on Boolean algebra* (with S. Skyum), J. Assoc. Comput. Mach. **32** (1985), 484–502.
20. *A logarithmic time sort for linear size networks* (with J. H. Reif), J. Assoc. Comput. Mach. **34** (1987), 60–76.
21. *NP is as easy as detecting unique solutions* (with V. V. Vazirani), Theoret. Comput. Sci. (to appear).

INSTITUT FÜR ANGEWANDTE MATHEMATIK DER UNIVERSITÄT ZÜRICH, 8032
ZÜRICH, SWITZERLAND

INVITED ONE-HOUR ADDRESSES
AT THE PLENARY SESSIONS

Underlying Concepts in the Proof of the Bieberbach Conjecture

LOUIS DE BRANGES

Most mathematical events are exciting only to a few experts who have the necessary knowledge to understand and appreciate what has occurred. The proof of the Bieberbach conjecture has been a welcome exception to this usual pattern. The result has found a wider, but also a less homogeneous, audience. For this reason there is need for still another account of the discovery.

A calendar of events. This discussion begins with the Riemann mapping theorem because everyone knows it: Every proper simply connected subregion of the complex plane is the image of the unit disk under a Riemann mapping function, a power series which converges in the unit disk and represents a function with distinct values at distinct points of the disk.

Riemann stated but did not find an acceptable proof of the Riemann mapping theorem. It is proved by an approximation argument which requires estimates of Riemann mapping functions. Any Riemann mapping function can be estimated in the unit disk from the value of the function and its derivative at the origin. The Bieberbach conjecture arises in the search for best possible estimates.

A Riemann mapping function will be normalized so that the constant coefficient is zero and the coefficient of z is positive. The mapping thus has a fixed point at the origin. The positivity of the derivative at the origin allows the mapping function to be uniquely determined by the region onto which it maps the unit disk.

An example of a Riemann mapping function is the Koebe function

$$z/(1 - \omega z)^2 = z + 2\omega z^2 + 3\omega^2 z^3 + \cdots,$$

where ω is a constant of absolute value one. It maps the unit disk onto a region which is obtained from the complex plane on deleting a radial slit starting at distance $1/4$ from the origin.

The Bieberbach conjecture [4] states that the inequality

$$|a_n| \leq na_1$$

Research supported by the National Science Foundation.

holds for every positive integer n when $f(z) = \sum a_n z^n$ is a Riemann mapping function. The conjecture also states that equality holds for some $n > 1$ only when $f(z)$ is a constant multiple of the Koebe function.

Bieberbach proved the Bieberbach conjecture only for the second coefficient. Despite the meager evidence in its favor, the conjecture attracted the attention of leading mathematicians. A memorable event was the proof of the Bieberbach conjecture for the third coefficient by K. Löwner [25] in 1923. The solution contains a general method which applies in principle to all coefficients. But the resulting calculations are very difficult. When a proof of the Bieberbach conjecture for the fourth coefficient was found in 1954 by P. R. Garabedian and M. Schiffer [19], another method was used.

The Bieberbach conjecture was actively pursued as a research aim in the following decade. A proof of the Bieberbach conjecture for the sixth coefficient was obtained in 1968 by M. Ozawa [34] and by R. N. Pederson [35]. In 1972, Pederson and Schiffer [36] succeeded in proving the Bieberbach conjecture for the fifth coefficient.

The same methods become very difficult with larger coefficients, and more than ten years passed without another case of the Bieberbach conjecture being verified. What then happened was that the author obtained a proof of the Bieberbach conjecture for all coefficients.

The proof was concluded with the help of Walter Gautschi [20] in February 1984. A curious situation then arose in that no one could at first be found to confirm the argument. But the author was scheduled for a visit to the Steklov Mathematical Institute in Leningrad in April, May, and June of 1984 under the exchange agreement between the National Academy of Sciences and the Academy of Sciences of the USSR. The correctness of the proof was confirmed by the Leningrad Seminar in Geometric Function Theory during the course of the visit.

The verification of the proof has been reported by three participants. One is the author [10]. The other two are the senior members of the Leningrad Seminar in Geometric Function Theory, G. V. Kuz'mina [18] and I. M. Milin [27].

The author expresses his thanks to the National Academy of Sciences and to the Academy of Sciences of the USSR for the exchange agreement which made the visit possible.

Interpolation theory. The reason why the Bieberbach conjecture is interesting is because of interpolation theory. All mathematical calculations, whether or not they are machine assisted, are necessarily finite. Therefore infinite processes need to be approximated by finite ones. Interpolation theory does this in the context of analytic function theory.

A function which is analytic in the unit disk is represented by a power series. The values of the function, especially on the boundary of the disk, can be very complicated. But the coefficients of the power series provide a natural means of finite approximation.

Some properties of the function are however not easily seen from the coefficients. A good example is boundedness. How should the coefficients of a power series be chosen so that the resulting function is bounded by one in the unit disk? The answer to the question was given in 1911 by C. Carathéodory and L. Fejér [17].

Boundedness is an important property of analytic functions because of the relation between the factorization of bounded analytic functions and invariant subspaces, a relation which first made its appearance in fundamental papers of A. Beurling [3] and M. S. Livšic and V. P. Potapov [24]. The interpolation theory of bounded analytic functions was developed within an invariant subspace context by D. Sarason [40]. It was translated into the Sz.-Nagy-Foiaş formulation of invariant subspace theory [45] by Sz.-Nagy and Foiaş [46], where it was generalized to power series whose coefficients are operators on a Hilbert space.

A related but different invariant subspace theory is due to James Rovnyak and the author [16]. This theory offers a natural context in which to formulate and prove the Carathéodory-Fejér result [7]. Some preliminaries from the theory of square summable power series [15] are needed.

The notation $\mathcal{C}(z)$ is used for the Hilbert space of square summable power series $f(z) = \sum a_n z^n$,

$$\|f(z)\|_{\mathcal{C}(z)}^2 = \sum |a_n|^2.$$

A power series $B(z)$ represents a function which is bounded by one in the unit disk if, and only if, multiplication by $B(z)$ is a contractive transformation in $\mathcal{C}(z)$.

Direct use of the contractive property of the transformation requires a concept called complementation. This is a generalization of orthogonality which was introduced by James Rovnyak and the author [16].

If a Hilbert space \mathcal{P} is contained contractively in a Hilbert space \mathcal{H} , then a unique Hilbert space \mathcal{Q} exists, which is contained contractively in \mathcal{H} and which has these properties: The inequality

$$\|c\|_{\mathcal{H}}^2 \leq \|a\|_{\mathcal{P}}^2 + \|b\|_{\mathcal{Q}}^2$$

holds whenever $c = a + b$ with a in \mathcal{P} and b in \mathcal{Q} . Every element c of \mathcal{H} admits a decomposition for which equality holds.

The Hilbert space \mathcal{Q} is called the complementary space to \mathcal{P} in \mathcal{H} . Minimal decomposition is unique. The element a of \mathcal{P} is obtained from c under the adjoint of the inclusion of \mathcal{P} in \mathcal{H} . The element b of \mathcal{Q} is obtained from c under the adjoint of the inclusion of \mathcal{Q} in \mathcal{H} .

If $B(z)$ is a power series which represents a function which is bounded by one in the unit disk, then the range $\mathcal{M}(B)$ of multiplication by $B(z)$ in $\mathcal{C}(z)$ is considered a Hilbert space in the unique norm such that multiplication by $B(z)$ is a partial isometry of $\mathcal{C}(z)$ onto $\mathcal{M}(B)$. The complementary space $\mathcal{H}(B)$ to $\mathcal{M}(B)$ in $\mathcal{C}(z)$ is invariant under the difference-quotient transformation, taking

$f(z)$ into $[f(z) - f(0)]/z$. The inequality for difference-quotients

$$\|[f(z) - f(0)]/z\|_{\mathcal{H}(B)}^2 \leq \|f(z)\|_{\mathcal{H}(B)}^2 - |f(0)|^2$$

is satisfied.

If $B'(z) = zB(z)$, then a space $\mathcal{H}(B')$ exists and the space $\mathcal{H}(B)$ is contained contractively in the space $\mathcal{H}(B')$. The complementary space $\mathcal{B}(B)$ to $\mathcal{H}(B)$ in $\mathcal{H}(B')$ is a Hilbert space which has dimension zero or one. Consider the complex numbers \mathcal{C} as a Hilbert space with absolute value as norm. Then multiplication by $B(z)$ is a partial isometry of \mathcal{C} onto $\mathcal{B}(B)$.

The Carathéodory-Fejér theorem can be formulated in a localization of the above structure. If r is a given nonnegative integer, define power series $f(z)$ and $g(z)$ to be r -equivalent if the coefficient of z^n in $f(z)$ is equal to the coefficient of z^n in $g(z)$ for $n = 0, \dots, r-1$. Define $\mathcal{C}_r(z)$ to be the finite-dimensional Hilbert space of r -equivalence classes of power series $f(z) = \sum a_n z^n$,

$$\|f(z)\|_{\mathcal{C}_r(z)}^2 = \sum_{n=0}^{r-1} |a_n|^2.$$

If $B(z)$ is a power series which represents a function which is bounded by one in the unit disk, then multiplication by $B(z)$ is contractive in $\mathcal{C}_r(z)$.

A modern formulation of the Carathéodory-Fejér theorem states that this property characterizes the r -equivalence class of a power series which represents a function which is bounded by one in the unit disk. A power series $B(z)$ such that multiplication by $B(z)$ is contractive in $\mathcal{C}_r(z)$ is r -equivalent to a power series which is bounded by one in the unit disk.

The proof of the theorem proceeds by an inductive construction of the coefficients. Assume that r is positive and that $A(z)$ is a power series such that multiplication of $A(z)$ is contractive in $\mathcal{C}_{r-1}(z)$. Then a power series $B(z)$ exists, which is r -equivalent to $A(z)$, such that multiplication of $B(z)$ is contractive in $\mathcal{C}_r(z)$.

The inductive step is made by working out the relationship between associated spaces. Define $\mathcal{M}_{r-1}(A)$ to be the range of multiplication by $A(z)$ in $\mathcal{C}_{r-1}(z)$. It is considered a Hilbert space in the unique norm such that multiplication by $A(z)$ is a partial isometry of $\mathcal{C}_{r-1}(z)$ onto $\mathcal{M}_{r-1}(A)$. Then the space $\mathcal{M}_{r-1}(A)$ is contained contractively in the space $\mathcal{C}_{r-1}(z)$. Define $\mathcal{H}_{r-1}(A)$ to be the complementary space to $\mathcal{M}_{r-1}(A)$ in $\mathcal{C}_{r-1}(z)$.

Let $A'(z) = zA(z)$. Then multiplication by $A'(z)$ is a contractive transformation in $\mathcal{C}_r(z)$. Define $\mathcal{M}_r(A')$ to be the range of multiplication by $A'(z)$ in $\mathcal{C}_r(z)$. Consider $\mathcal{M}_r(A')$ in the unique norm such that multiplication by $A'(z)$ is a partial isometry of $\mathcal{C}_r(z)$ onto $\mathcal{M}_r(A')$. The $\mathcal{M}_r(A')$ is contained contractively in $\mathcal{C}_r(z)$. Define $\mathcal{H}_r(A')$ to be the complementary space to $\mathcal{M}_r(A')$ in $\mathcal{C}_r(z)$. Then $\mathcal{H}_r(A')$ is the set of r -equivalence classes of power series $f(z)$ such that $[f(z) - f(0)]/z$ belongs to $\mathcal{M}_{r-1}(A)$. The identity

$$\|[f(z) - f(0)]/z\|_{\mathcal{M}_{r-1}(A)}^2 = \|f(z)\|_{\mathcal{M}_r(A')}^2 - |f(0)|^2$$

holds for every element $f(z)$ of $\mathcal{M}_r(A')$.

The construction of the desired power series $B(z)$ is made by constructing any Hilbert space \mathcal{H}_r that is contained contractively in $\mathcal{H}_r(A')$ such that the transformation which takes every element of \mathcal{H}_r into its $(r-1)$ -equivalence class is a partial isometry of \mathcal{H}_r onto $\mathcal{H}_{r-1}(A)$. Let \mathcal{B}_r be the complementary space to \mathcal{H}_r in $\mathcal{H}_r(A')$. Then \mathcal{B}_r contains a power series $B(z)$ of norm zero or one which is $(r-1)$ -equivalent to $A(z)$. Multiplication by $B(z)$ is contractive in $\mathcal{C}_r(z)$.

This formulation of the Carathéodory-Fejér theorem also applies to power series whose coefficients are operators on a Hilbert space. It can be used to obtain information about the relation between factorization and invariant subspaces [14].

The relationship between the Sz.-Nagy-Foias invariant subspace theory and the present theory of square summable series is discussed by N. K. Nikol'skii and V. I. Vasyunin [33] and by D. Sarason [41]. The relation can also be formulated in the language of systems theory [12].

A proof of the Carathéodory-Fejér theorem outside of the present theory of square summable power series is given by M. Rosenblum and J. Rovnyak [39]. See also N. K. Nikol'skii [32] for another modern treatment of interpolation theory.

An indefinite generalization of the Hilbert space concept was introduced into interpolation theory by R. Nevanlinna [28, 29, 30, 31]. A generalization of complementation theory applies in these spaces [13].

An alternative to the Carathéodory-Fejér method for the interpolation theory of bounded functions is given by the Schur algorithm. D. Alpay and H. Dym [1] show that similar Hilbert spaces of analytic functions are the outcome of the theory.

Bounded Riemann mapping functions. The interpolation problem for bounded analytic functions acquires added interest when the function is required to be a Riemann mapping function. This condition is interesting because it is compatible with the interpolation problem for bounded functions.

A Hilbert space which is relevant to the study of bounded Riemann mapping functions is the set of power series

$$f(z) = \sum_{n=1}^{\infty} a_n z^n$$

with constant coefficient zero which have finite Dirichlet norm,

$$\|f(z)\|_{\mathcal{G}}^2 = \sum_{n=1}^{\infty} n|a_n|^2.$$

The elements of the space are convergent power series in the unit disk. The Dirichlet norm is important because it is computable from the integral square of the derivative in the unit disk,

$$\|f(z)\|_{\mathcal{G}}^2 = \frac{1}{\pi} \iint_{\Delta} |f'(z)|^2 dx dy.$$

If $B(z)$ is a normalized Riemann mapping function for a subregion of the unit disk, then

$$\begin{aligned}\|f(B(z))\|_{\mathcal{G}}^2 &= \frac{1}{\pi} \iint_{\Delta} |f'(B(z))|^2 |B'(z)|^2 dx dy \\ &= \frac{1}{\pi} \iint_{B(\Delta)} |f'(z)|^2 dx dy\end{aligned}$$

by the change of variable theorem for multiple integrals because $|B'(z)|^2$ coincides with the Jacobian of the mapping by the Cauchy-Riemann equations. Since the image $B(\Delta)$ of the unit disk is contained in the unit disk Δ by hypothesis, the inequality

$$\iint_{B(\Delta)} |f'(z)|^2 dx dy \leq \iint_{\Delta} |f'(z)|^2 dx dy$$

is satisfied. It follows that the inequality

$$\|f(B(z))\|_{\mathcal{G}}^2 \leq \|f(z)\|_{\mathcal{G}}^2$$

holds for every element $f(z)$ of the space \mathcal{G} .

The inequality states that $B(z)$ -substitution, the transformation which takes $f(z)$ into $f(B(z))$, is contractive in the Dirichlet norm. This observation leads to several different preliminaries to the proof of the Bieberbach conjecture [5, 6, 8], which will now be presented.

Consider the range $\mathcal{N}(B)$ of $B(z)$ -substitution as a Hilbert space in the unique norm such that $B(z)$ -substitution is an isometry of \mathcal{G} onto $\mathcal{N}(B)$. Then the space $\mathcal{N}(B)$ is contained contractively in \mathcal{G} . Define $\mathcal{G}(B)$ to be the complementary space to $\mathcal{N}(B)$ in \mathcal{G} . Then $\mathcal{G}(B)$ is a Hilbert space which has a computable reproducing kernel function.

Indeed the reproducing kernel function of the space \mathcal{G} is

$$\log \frac{1}{1 - z\bar{w}} = z\bar{w} + \frac{1}{2}(z\bar{w})^2 + \frac{1}{3}(z\bar{w})^3 + \cdots,$$

the reproducing kernel function of $\mathcal{N}(B)$ is

$$\log \frac{1}{1 - B(z)\overline{B(w)}},$$

and the reproducing kernel function of $\mathcal{G}(B)$ is

$$\log \frac{1 - B(z)\overline{B(w)}}{1 - z\bar{w}} = \log \frac{1}{1 - z\bar{w}} - \log \frac{1}{1 - B(z)\overline{B(w)}}.$$

This result is to be compared with the calculation of the reproducing kernel function of the space $\mathcal{M}(B)$. The reproducing kernel function of $\mathcal{C}(z)$ is

$$\frac{1}{1 - z\bar{w}},$$

the reproducing kernel function of $\mathcal{M}(B)$ is

$$\frac{B(z)\overline{B(w)}}{1 - z\bar{w}},$$

and the reproducing kernel function of $\mathcal{H}(B)$ is

$$\frac{1 - B(z)\overline{B}(w)}{1 - z\bar{w}} = \frac{1}{1 - z\bar{w}} - \frac{B(z)\overline{B}(w)}{1 - z\bar{w}}.$$

Note that the reproducing kernel function of the space $\mathcal{H}(B)$ is the exponential of the reproducing kernel function of the space $\mathcal{G}(B)$. An inequality is now a consequence of the theory of reproducing kernel functions. If $f(z)$ belongs to $\mathcal{G}(B)$, then $\exp f(z)$ belongs to $\mathcal{H}(B)$ and the inequality

$$\|\exp f(z)\|_{\mathcal{H}(B)}^2 \leq \exp \|f(z)\|_{\mathcal{G}(B)}^2$$

is satisfied. Equality holds when $f(z)$ is the reproducing kernel function of the space $\mathcal{G}(B)$ corresponding to some point of the unit disk. If $f(z)$ is an element of $\mathcal{G}(B)$ for which equality holds, then the identity

$$\langle \exp f(z), \exp g(z) \rangle_{\mathcal{H}(B)} = \exp \langle f(z), g(z) \rangle_{\mathcal{G}(B)}$$

holds for every element $g(z)$ of $\mathcal{G}(B)$.

These relations support the appropriateness of the Dirichlet norm for the estimation theory of Riemann mapping functions. Now a natural generalization of the Dirichlet norm exists for Laurent series. The quadratic forms which then arise are indefinite. But similar methods still apply. In fact it is possible to introduce generalizations of power series in which the exponents are not integers.

If ν is a given real number, define \mathcal{G}^ν to be the set of generalized power series

$$f(z) = \sum_{n=1}^{\infty} a_n z^{\nu+n}$$

such that

$$\langle f(z), f(z) \rangle_{\mathcal{G}^\nu} = \sum_{n=1}^{\infty} (\nu + n) |a_n|^2$$

is finite. Note that only a finite number of squares appear with negative coefficients. The infinite sum is therefore meaningful.

If $B(z)$ is a normalized Riemann mapping function for a subregion of the unit disk, then the substituted series $f(B(z))$ has a natural interpretation as an element of \mathcal{G}^ν whenever $f(z)$ belongs to \mathcal{G}^ν , and the inequality

$$\langle f(B(z)), f(B(z)) \rangle_{\mathcal{G}^\nu} \leq \langle f(z), f(z) \rangle_{\mathcal{G}^\nu}$$

is satisfied.

The inequality has a long history. When $\nu = -2$, the inequality is a strengthening of the area theorem, discovered by T. H. Gronwall [22] and used by Bieberbach [4] to prove the Bieberbach conjecture for the second coefficient. The area theorem is a limiting case of the present inequality which applies to unbounded Riemann mapping functions.

In 1939, H. Grunsky [23] obtained a generalization of the area theorem. Again the result is stated for unbounded normalized Riemann mapping functions. It is a limiting case of the present inequality when ν is a negative integer. The Grunsky

inequality was instrumental in the proof of the Bieberbach conjecture for the fourth, fifth, and sixth coefficients. Another reason for interest in the Grunsky inequality is that it characterizes normalized Riemann mapping functions.

A postwar generation of analysts found a stronger form of the inequality, which applies to normalized Riemann mapping functions for subregions of the unit disk. In fact, the inequality characterizes these functions. This information has been used by Tammi [47, 48] to characterize the first four coefficients of a bounded Riemann mapping function. The inequality is equivalent to the above-stated inequality for negative integers ν , but it is not stated in that form. The observation that normalized Riemann mapping functions for subregions of the unit disk are characterized by contractive properties of substitution transformations with respect to indefinite scalar products is due to the author [15].

These results make use of the space \mathcal{G}^ν only for integer values of ν . The discovery that a variant of the area theorem holds for noninteger values of ν is due to Prawitz [37].

Adjoints are another consideration in the proof of the Bieberbach conjecture. A bounded linear transformation of a Hilbert space into itself has an adjoint. If the transformation is in some way computable and if the Hilbert space norm is well related to the transformation, then the adjoint can be expected to be computable in an analogous way.

These expectations are fulfilled for substitution transformations considered with the Dirichlet norm and its indefinite generalizations. Consider spaces \mathcal{G}^μ and \mathcal{G}^ν such that $\mu + \nu + 1$ is a negative integer. If $g(z)$ belongs to \mathcal{G}^μ , define $P_\nu g(1/z)$ to be the unique element of \mathcal{G}^ν such that $f(z) = \sum a_n z^{\nu+n}$ and $g(z) = \sum b_n z^{\mu+n}$, where $a_m = b_n$ whenever $\mu + \nu + m + n = 0$ for positive integers m and n and such that $a_m = 0$ otherwise. If $f(z)$ belongs to \mathcal{G}^ν and if $g(z)$ belongs to \mathcal{G}^μ , then the identity

$$\langle f(z), P_\nu g(1/z) \rangle_{\mathcal{G}^\nu} = -\langle P_\mu f(1/z), g(z) \rangle_{\mathcal{G}^\mu}$$

is satisfied.

Assume that $B(z) = \sum B_n z^n$ is a normalized Riemann mapping function for a subregion of the unit disk. Then a normalized Riemann mapping function $B^*(z)$ for a subregion of the unit disk is defined by $B^*(z) = \sum \bar{B}_n z^n$. The region onto which $B^*(z)$ maps the unit disk is the reflection in the real axis of the region onto which $B(z)$ maps the unit disk.

If $f(z)$ is an element of \mathcal{G}^ν and if $g(z)$ is an element of \mathcal{G}^μ , then the identity

$$\langle f(z), P_\nu g(1/z) \rangle_{\mathcal{G}^\nu} = \langle f(B(z)), P_\nu g(B^*(1/z)) \rangle_{\mathcal{G}^\nu}$$

is satisfied. The identity is proved using Cauchy's formula.

As a result of the identity, $B(z)$ -substitution has a computable adjoint, at least on those elements of \mathcal{G}^ν of the form $P_\nu g(B^*(1/z))$ where $g(z)$ belongs to \mathcal{G}^μ . The action of the adjoint of $B(z)$ -substitution on such an element of \mathcal{G}^ν is $P_\nu g(1/z)$. Thus the adjoint of $B(z)$ -substitution in \mathcal{G}^ν corresponds to the inverse of $B^*(z)$ -substitution in \mathcal{G}^μ . By the arbitrariness of μ , this result computes the adjoint of $B(z)$ -substitution on a dense set of elements of \mathcal{G}^ν .

These considerations allow a new synthesis of the theory of Riemann mapping functions [15]. It is interesting to compare the approach taken here with that of I. Schur [42, 43, 44], who was a founder of operator theory and who made appropriate calculations with formal power series and indefinite scalar products. But Schur did not have available the methods of the theory of square summable power series for describing what is happening. This conceptual advantage was instrumental in the proof of the Bieberbach conjecture.

Localization of estimates. The mechanism which has been constructed for estimating the coefficients of Riemann mapping functions suffers from a deficiency, namely, that the estimates make use of all the coefficients of a power series. By analogy with the Carathéodory-Fejér theorem, it is interesting to find related estimates which depend only on the first r coefficients of a power series for some given positive integer r . Truncation does not however now produce best possible estimates. The proof of the Bieberbach conjecture makes use of a more delicate localization technique.

The source of the new estimates is the theory with which K. Löwner proved the Bieberbach conjecture for the third coefficient.

Subordination is a key idea in the Löwner theory. A power series $f(z)$ with constant coefficient zero is said to be subordinate to a power series $g(z)$ with constant coefficient zero if $f(z) = g(B(z))$ for a power series $B(z)$ with constant coefficient zero which represents a function which is bounded by one in the unit disk. When $f(z)$ and $g(z)$ are Riemann mapping functions, an equivalent condition is that the region onto which $f(z)$ maps the unit disk is contained in the region onto which $g(z)$ maps the unit disk. The series $B(z)$ so obtained is then a Riemann mapping function, and it is normalized if $f(z)$ and $g(z)$ are normalized.

A Löwner family of Riemann mapping functions is a maximal family of normalized Riemann mapping functions which is totally ordered in the sense of subordination. Such a family has a natural parametrization. The parameter is the coefficient of z in the power series. All positive numbers appear as parameters.

Löwner made the interesting discovery that a Löwner family of Riemann mapping functions $F(t, z)$ satisfies a differential equation,

$$t \frac{\partial}{\partial t} F(t, z) = \varphi(t, z) z \frac{\partial}{\partial z} F(t, z).$$

The equation states that a time derivative is proportional to a space derivative. The constant of proportionality $\varphi(t, z)$ is a Herglotz function, a power series with constant coefficient one which represents a function with positive real part in the unit disk. The family of Herglotz functions is measurable in the sense that the coefficient of z^n in $\varphi(t, z)$ is a measurable function of t for every nonnegative integer n . The time derivative in the Löwner equation is taken in the sense of absolute continuity for coefficients.

A converse result is also true. Assume that a measurable family $\varphi(t, z)$ of Herglotz functions is given. Then a unique Löwner family of Riemann mapping functions exists which has the given family $\varphi(t, z)$ as coefficient function in the Löwner equation.

In principle all information about Riemann mapping functions is obtainable from calculations with the Löwner equation. But these calculations become very difficult when a precise goal like the Bieberbach conjecture is set. A convenient form of coding is needed to propagate estimates in time. Quadratic forms which are well behaved with respect to the propagation are the natural answer. The propagation should either preserve the quadratic forms or be dissipative with respect to them.

The Bieberbach conjecture is not itself a quadratic estimate, but a related quadratic estimate was conjectured in 1936 by M. S. Robertson [38]. The statement of the conjecture makes use of a construction for powers of Riemann mapping functions which can be made for any real exponent ν . If $F(z)$ is a normalized Riemann mapping function, then the formal expression $F(z)^\nu$ can be expanded as a series whose first term is $f'(0)^\nu z^\nu$. A convenient way to write the remaining terms in the expression is

$$\frac{F(z)^\nu - F'(0)^\nu z^\nu}{\nu} = \sum_{n=1}^{\infty} a_n z^{\nu+n}.$$

In the limiting case $\nu = 0$, the expression on the left is interpreted by continuity as

$$\log \frac{F(z)}{zF'(0)}.$$

When ν is not an integer, a meaning is given to the expansion by dividing each side of the equation by z^ν and choosing the unique ν th power of $F(z)/z$, which is a power series with positive constant coefficient.

In this notation the Robertson conjecture states that the inequality

$$|a_1|^2 + \cdots + |a_r|^2 \leq 4rF'(0)$$

holds for every positive integer r when $\nu = \frac{1}{2}$. It is easily seen that equality holds when $F(z)$ is a constant multiple of the Koebe function, and Robertson conjectured that equality holds only in that case. An elementary estimate shows that the inequality implies the Bieberbach conjecture for the $(r+1)$ st coefficient. Robertson verified the cases of the Robertson conjecture corresponding to the Bieberbach conjecture for the second and third coefficients. Thus he showed that all of the then available evidence for the Bieberbach conjecture was matched by evidence for the stronger Robertson conjecture. And his method of proof was the Löwner equation. In retrospect this combination of quadratic forms with the Löwner equation is seen as a remarkable penetration into the estimation theory of Riemann mapping functions. More information related to the Bieberbach conjecture would have been obtained at that time if the method had been further pursued. But the necessary effort was not made because there was then

insufficient belief in the Bieberbach conjecture. When new evidence in its favor was later obtained, attention was drawn away from the Löwner equation because the new results were obtained by other methods.

A return to the quadratic ideas of Robertson was made in 1965 when Lebedev and Milin [26] observed that the inequality

$$\sum_{n=1}^r n(r+1-n)|a_n|^2 \leq 4(r+1) \sum_{n=1}^r 1/(n+1)$$

in the case $\nu = 0$ implies the Bieberbach conjecture for the $(r+1)$ st coefficient. In fact they show that the inequality implies the corresponding case of the Robertson conjecture. The Milin conjecture states that the inequality is always satisfied, and that equality holds only for a constant multiple of the Koebe function.

The Robertson and Milin conjectures suggest the construction of a new space, whose elements are (equivalence classes of) generalized power series $f(z) = \sum_{n=1}^{\infty} a_n z^{\nu+n}$. Assume that σ_n is a given real-valued function of positive integers n which is eventually zero. Define

$$\langle f(z), f(z) \rangle_{g_{\nu}^{\sigma}} = \sum_{n=1}^{\infty} (\nu+n) \sigma_n |a_n|^2.$$

Equivalence of two such generalized power series $f(z)$ and $g(z)$ here means that the coefficient of $z^{\nu+n}$ in $f(z)$ is equal to the coefficient of $z^{\nu+n}$ in $g(z)$ when $(\nu+n)\sigma_n$ is not zero.

The Robertson and Milin conjectures are now estimates of the same form. The problem is to estimate the expression

$$\left\langle \frac{F(z)^{\nu} - F'(0)^{\nu} z^{\nu}}{\nu}, \frac{F(z)^{\nu} - F'(0)^{\nu} z^{\nu}}{\nu} \right\rangle_{g_{\nu}^{\sigma}},$$

where $F(z)$ is a normalized Riemann mapping function. The question is whether the expression is maximized, for a given choice of $F'(0)$, with $F(z)$ equal to a constant multiple of the Koebe function.

The answer can be expected to depend on the choice of σ , and two-dimensional examples show that this is indeed the case. The question then is what weights make the optimization problem have the anticipated solution.

An answer is suggested by the above characterization of bounded Riemann mapping functions. The estimation theory can be expected to be a limiting case of an estimation theory applying to Riemann mapping functions for subregions of the unit disk. The feature of σ which makes it have the desired property is the contractive property of related substitution transformations.

The answer is further complicated by an unexpected twist which has no analogue in the Carathéodory-Fejér theory. This is that the weights σ must be allowed to change during the substitution process. The contractive property of the transformation is obtained in passing from one space $\mathcal{G}_{\sigma}^{\nu}$ to another. The

need to make such weight changes and the precise way in which they should be made is one of the more subtle aspects of the proof of the Bieberbach conjecture.

Assume that 2ν is not a negative integer. A family of spaces $\mathcal{G}_{\sigma(t)}^\nu$, $t \geq 1$, is said to be admissible if $(\nu + n)\sigma_n(t)$ is an absolutely continuous function of t which is nonincreasing when $2\nu + n > 0$ and nondecreasing when $2\nu + n < 0$, and if the differential equation

$$\sigma_n(t) + \frac{t\sigma'_n(t)}{2\nu + n} = \sigma_{n+1}(t) - \frac{t\sigma'_{n+1}(t)}{n+1}$$

holds for every positive integer n . Solutions will be considered in which $\sigma_n(t)$ is eventually identically zero.

These conditions are carefully chosen to imply the inequality

$$\langle f(B(z)), f(B(z)) \rangle_{\mathcal{G}_{\sigma(a)}^\nu} \leq \langle f(z), f(z) \rangle_{\mathcal{G}_{\sigma(b)}^\nu}$$

for every element $f(z)$ of $\mathcal{G}_{\sigma(b)}^\nu$, when $B(z)$ is a normalized Riemann mapping function for a subregion of the unit disk and $1 \leq a = bB'(0)$. The proof of the inequality from the stated hypotheses is a straightforward application of the Löwner differential equation [11, 15].

Once the result has been obtained, the question arises whether it is possible to obtain an estimate of the desired expression

$$\left\langle \frac{F(z)^\nu - F'(0)^\nu z^\nu}{\nu}, \frac{F(z)^\nu - F'(0)^\nu z^\nu}{\nu} \right\rangle_{\mathcal{G}_{\sigma(a)}^\nu}$$

when $F(z)$ is a normalized Riemann mapping function. The desired maximum of the expression, obtained when $F(z)$ is a constant multiple of the Koebe function, is

$$4F'(0)^{2\nu} \sum_{n=1}^{\infty} \frac{\Gamma(2\nu + n)^2}{\Gamma(2\nu + 1)^2 \Gamma(n + 1)^2} (\nu + n) \sigma_n(a).$$

It is sufficient to obtain the estimate for bounded functions. A stronger estimate can be expected in that case. Indeed another application of the Löwner equation shows that the inequality

$$\begin{aligned} & \left\langle \frac{B(z)^\nu - B'(0)^\nu z^\nu}{\nu} + f(B(z)), \frac{B(z)^\nu - B'(0)^\nu z^\nu}{\nu} + f(B(z)) \right\rangle_{\mathcal{G}_{\sigma(a)}^\nu} \\ & \leq \langle f(z), f(z) \rangle_{\mathcal{G}_{\sigma(b)}^\nu} + \sum_{n=1}^{\infty} \frac{\Gamma(2\nu + n)^2}{\Gamma(2\nu + 1)^2 \Gamma(n + 1)^2} (\nu + n) [a^{2\nu} \sigma_n(a) - b^{2\nu} \sigma_n(b)] \end{aligned}$$

holds for every element $f(z)$ of $\mathcal{G}_{\sigma(b)}^\nu$ when $B(z)$ is a normalized Riemann mapping function for a subregion of the unit disk and $1 \leq a = bB'(0)$. Furthermore the proof shows that equality holds if, and only if, a complex number ω of absolute value one exists such that

$$\frac{B(z)}{(1 + \omega B(z))^2} = \frac{B'(0)z}{(1 + \omega z)^2}$$

and such that the coefficient of $z^{\nu+n}$ in $f(z)$ is equal to the coefficient of $z^{\nu+n}$ in

$$\frac{z^\nu}{\nu(1+\omega z)^{2\nu}} - \frac{z^\nu}{\nu}$$

when $\sigma_n(t)$ is not identically zero.

Thus the cases of equality are related to the Koebe function. The desired estimate of unbounded Riemann mapping functions is obtained in the limit of large b . The cases of equality do occur only when the Riemann mapping function is a constant multiple of the Koebe function.

These considerations supply the desired estimates of Riemann mapping functions at the cost of information about the existence of monotone solutions of differential equations.

A solution of the system of differential equations for the coefficient functions is obtained of the form

$$\begin{aligned} \sigma_n(t) = & \Delta_{2\nu+n}(t) + \frac{(2\nu+n)(2\nu+2n+2)}{(-1)(n+1)} \Delta_{2\nu+n+1}(t) \\ & + \frac{(2\nu+n)(2\nu+n+1)(2\nu+2n+3)(2\nu+2n+4)}{(-1)(-2)(n+1)(n+2)} \Delta_{2\nu+n+2}(t) + \dots, \end{aligned}$$

where $\Delta_{2\nu+n}(t)$ is a solution of the elementary differential equation

$$t\Delta'_{2\nu+n}(t) = -(2\nu+n)\Delta_{2\nu+n}(t).$$

The solution is

$$\Delta_{2\nu+n}(t) = \Delta_{2\nu+n}(1)t^{-2\nu-n}.$$

Examples of solutions are given in terms of hypergeometric series. The notation

$$F(a, b; c; z) = 1 + \frac{ab}{1 \cdot c}z + \frac{a(a+1)b(b+1)}{1 \cdot 2c(c+1)}z^2 + \dots$$

is used for the hypergeometric series. A generalization of the hypergeometric series is

$$F(a, b, c; d, e; z) = 1 + \frac{abc}{1 \cdot de}z + \frac{a(a+1)b(b+1)c(c+1)}{1 \cdot 2d(d+1)e(e+1)}z^2 + \dots$$

An identity due to Clausen states that

$$F(a, b; c; z)^2 = F(2a, 2b, a+b; c, 2c-1; z)$$

when $a+b+\frac{1}{2}=c$.

A nonnegative parameter λ will be used in constructing solutions of the differential equations for the coefficient functions. Let r be a given positive integer. Choose

$$\Delta_{2\nu+n}(1) = \frac{4^{-n}\Gamma(n+1)\Gamma(2\nu+2\lambda+r+n+1)}{\Gamma(\nu+n+1)\Gamma(\nu+\lambda+n+1)\Gamma(2\nu+n+1)\Gamma(r+1-n)}$$

for $n = 1, \dots, r$, and $\Delta_{2\nu+n}(1) = 0$ for $n > r$. Then the identity

$$\begin{aligned} -t\sigma'_n(t) &= \frac{4^{-n}\Gamma(n+1)\Gamma(2\nu+2\lambda+r+n+1)}{\Gamma(\nu+n+1)\Gamma(\nu+\lambda+n+1)\Gamma(2\nu+n)\Gamma(r+1-n)} \\ &\quad \times F(n-r, 2\nu+2\lambda+r+n+1, \nu+n+\tfrac{1}{2}; \\ &\quad \nu+\lambda+n+1, 2\nu+2n+1; t^{-1})t^{-2\nu-n} \end{aligned}$$

holds for $n = 1, \dots, r$.

The polynomial

$$F(n-r, 2\nu+2\lambda+r+n+1, \nu+n+\tfrac{1}{2}; \nu+\lambda+n+1, 2\nu+2n+1, x)$$

is nonnegative for $x \leq 1$ by an inequality due to R. Askey and G. Gasper [2] if $\nu > -\frac{3}{2}$. In the case $\lambda = 0$, the inequality is due to Clausen's identity. When $\lambda > 0$, the inequality follows from an expansion

$$\begin{aligned} &\frac{\Gamma(2\nu+2\lambda+r+n+1)}{\Gamma(\nu+\lambda+n+1)\Gamma(r+1-n)} \\ &\quad \times F(n-r, 2\nu+2\lambda+r+n+1, \nu+n+\tfrac{1}{2}; \\ &\quad \nu+\lambda+n+1, 2\nu+2n+1; z) \\ &= \sum_{k=n}^r \frac{\Gamma(2\nu+k+n+1)}{\Gamma(\nu+n+1)\Gamma(k+1-n)} c_k \\ &\quad \times F(n-k, 2\nu+k+n+1, \nu+n+\tfrac{1}{2}; \nu+n+1, 2\nu+2n+1; z), \end{aligned}$$

where the numbers c_k are nonnegative. Indeed $c_n = 0$ where n has opposite parity to r , and

$$\begin{aligned} c_n &= 2^{2\lambda} \frac{\Gamma(\nu+\lambda+\tfrac{1}{2}r+\tfrac{1}{2}n+\tfrac{1}{2})(\nu+n+\tfrac{1}{2})}{\Gamma(\nu+\tfrac{1}{2}r+\tfrac{1}{2}n+\tfrac{3}{2})} \\ &\quad \times \frac{\Gamma(\lambda+\tfrac{1}{2}r-\tfrac{1}{2}n)}{\Gamma(\lambda)\Gamma(1+\tfrac{1}{2}r-\tfrac{1}{2}n)} \end{aligned}$$

when n has the same parity as r .

It follows that an admissible family of spaces is obtained when $\nu > -\frac{3}{2}$. The resulting estimates are rather complicated to state except in the case $\lambda = \frac{1}{2}$. Assume that σ_n is any real-valued function of positive integers n such that $\lim_{n \rightarrow \infty} \sigma_n = 0$ and such that

$$\rho_n = \frac{\Gamma(2\nu+n+1)}{\Gamma(n+1)} [\sigma_n - \sigma_{n+1}]$$

is nonnegative, nonincreasing, and has limit zero at infinity. If $F(z)$ is a normalized Riemann mapping function and if

$$\frac{F(z)^\nu - F'(0)^\nu z^\nu}{\nu} = \sum_{n=1}^{\infty} a_n z^{\nu+n},$$

then

$$\begin{aligned} & \sum_{n=1}^{\infty} (\nu + n) \sigma_n |a_n|^2 \\ & \leq 4F'(0)^{2\nu} \sum_{n=1}^{\infty} \frac{\Gamma(2\nu + n)^2}{\Gamma(2\nu + 1)^2 \Gamma(n + 1)^2} (\nu + n) \sigma_n. \end{aligned}$$

Equality holds with σ_n not identically zero if, and only if, $F(z)$ is a constant multiple of the Koebe function.

This result contains the Milin conjecture in the case $\nu = 0$, and it completes the proof of the Bieberbach conjecture.

Remarks. What has been presented here as a proof of the Bieberbach conjecture is in reality a general estimation theory for the coefficients of Riemann mapping functions. The Bieberbach conjecture is only significant as a historical marker. It is a test to be applied to an estimation theory to measure its strength. The result is interesting principally because it is difficult to obtain. The value of the work lies in the resulting estimation theory of Riemann mapping functions rather than in the Bieberbach conjecture itself.

The present proof of the Bieberbach conjecture proceeds through the verification of a conjecture of Milin, which implies the Robertson conjecture, which implies the Bieberbach conjecture. It is very likely that other routes will be found to prove the Bieberbach conjecture from the same estimation theory. The contractive properties of substitution transformations in the spaces \mathcal{G}_ν' where 2ν is a negative integer are likely to be useful for this purpose. These cases are more complicated because of singularities in the differential equations for coefficient functions. Another complication is due to the fact that the Askey-Gasper inequalities do not apply when $\nu < -\frac{3}{2}$. The estimation theory has to be considered afresh for these values of ν .

A time can be expected to come when interest in the Bieberbach conjecture itself will fade and more fundamental problems will again be considered. One of these is the interpolation problem for Riemann mapping functions. The problem is to characterize the first r coefficients a_1, \dots, a_r of a normalized Riemann mapping function $f(z) = \sum a_n z^n$. Computer calculations may be significant in testing the consequences of inequalities which are proposed for the coefficients.

The proof of the Bieberbach conjecture suggests that a less traditional problem may be even more fundamental. The problem is to characterize the first r coefficients a_1, \dots, a_r of a normalized Riemann mapping function $\sum a_n z^n$ for a subregion of the unit disk. If the coefficients of bounded functions can be characterized, then a characterization of coefficients of unbounded functions is deduced as a corollary.

The above estimation theory produces a large number of inequalities satisfied by the coefficients of a normalized Riemann mapping function for a subregion of the unit disk. The inequalities assert the contractive properties of substitution transformations in spaces of power series equipped with a scalar product. It is

natural to conjecture that the inequalities which apply to the first r coefficients of a bounded Riemann mapping function characterize these coefficients.

The proof of such a conjecture is difficult at present because there are too many inequalities. Some insight is needed as to what are the important inequalities. The choice of ν equal to a negative integer is likely to be important.

It may not be possible to determine, in any explicit way, all solutions of the differential equations for the coefficient functions with the desired monotonicity properties. But a characterization of coefficients will surely have to take all such solutions into account. That is a complicating feature of the analysis which has to be made.

An extension problem, similar to the one in the Carathéodory-Fejér theorem, can be expected as the main obstacle in coefficient characterization. Computer exploration may be necessary to determine whether the desired extendability is possible with proposed characterizing inequalities.

Finally the relation of the interpolation problem for bounded Riemann mapping functions to the Carathéodory-Fejér problem needs to be examined. Exponential relations are known to exist between the two interpolation theories before localization [15]. Exponential relations between the local theorems are needed. The Lebedev-Milin inequality [26], which is used in the proof of the Bieberbach conjecture, should be examined in that light. Perhaps a new proof of the inequality can be found which generalizes to related situations.

This paper was written while the author was on sabbatical leave from Purdue University as an Alexander von Humboldt fellow at the University of Heidelberg in March, April, May, June, and July of 1986. He thanks his hosts, Professors A. Dold and E. Freitag.

REFERENCES

1. D. Alpay and H. Dym, *An application of reproducing kernel Hilbert spaces to the Schur algorithm and rational J unitary factorization*, I. Schur Methods in Operator Theory and Signal Processing, Birkhäuser, Basel, 1986.
2. R. Askey and G. Gasper, *Inequalities for polynomials*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R.I., 1986, pp. 7–32.
3. A. Beurling, *On two problems concerning linear transformations in Hilbert space*, Acta Math. **81** (1949), 239–255.
4. L. Bieberbach, *Über die Koeffizienten derjenigen Potenzreihen welche eine schlichte Abbildung des Einheitskreises vermitteln*, Sitzungsbericht Preussische Akademie der Wissenschaften, 1916, pp. 940–955.
5. L. de Branges, *Coefficient estimates*, J. Math. Anal. Appl. **82** (1981), 420–450.
6. —, *Grunsky spaces of analytic functions*, Bull. Sci. Math. (2) **105** (1981), 401–416.
7. —, *The Carathéodory-Fejér extension theorem*, Integral Equations Operator Theory **5** (1982), 160–183.
8. —, *Löwner expansions*, J. Math. Anal. Appl. **100** (1984), 323–337.
9. —, *A proof of the Bieberbach conjecture*, Acta Math. **154** (1985), 137–152.
10. —, *The story of the verification of the Bieberbach conjecture*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 199–203.

11. —, *Powers of Riemann mapping functions*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 51–67.
12. —, *Unitary linear systems whose transfer functions are Riemann mapping functions*, Integral Equations Operator Theory (to appear).
13. —, *Complementation in Krein spaces*, unpublished manuscript, 1986.
14. —, *Krein spaces of analytic functions*, unpublished manuscript, 1986.
15. —, *Square summable power series*, Springer-Verlag, Heidelberg (to appear).
16. L. de Branges and J. Rovnyak, *Canonical models in quantum scattering theory*, Perturbation Theory and its Applications in Quantum Mechanics, Wiley, New York, 1966, pp. 295–391.
17. Carathéodory and L. Fejér, *Über den Zusammenhang der Extremen von Harmonischen Funktionen mit ihren Koeffizienten und über den Picard Landauschen Satz*, Rend. Circ. Mat. Palermo **32** (1911), 218–239.
18. O. M. Fomenko and G. V. Kuz'mina, *The last hundred days of the Bieberbach conjecture*, Math. Intelligencer **8** (1986), no. 1, 40–47.
19. P. R. Garabedian and M. Schiffer, *A proof of the Bieberbach conjecture for the fourth coefficient*, Arch. Rational Mech. Anal. **4** (1955), 427–465.
20. W. Gautschi, *Reminiscences of my involvement in de Branges' proof of the Bieberbach conjecture*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 205–211.
21. G. M. Goluzin, *Geometric theory of functions of a complex variable*, 2nd ed., "Nauka", Moscow, 1966; English transl., Transl. Math. Monographs, vol. 26, Amer. Math. Soc., Providence, R. I., 1969.
22. T. H. Gronwall, *Some remarks on conformal representation*, Ann. of Math. **16** (1914–1915), 72–76.
23. H. Grunsky, *Koeffizienten Bedingungen für schlicht abbildende meromorphe Funktionen*, Math. Z. **45** (1939), 26–61.
24. M. S. Livšic and V. P. Potatov, *The multiplication theorem for characteristic matrix functions*, Dokl. Akad. Nauk SSSR **62** (1950), 625–628. (Russian)
25. K. Löwner, *Untersuchungen über schlichte konforme Abbildungen des Einheitskreises*, Math. Ann. **89** (1923), 103–121.
26. I. M. Milin, *Univalent functions and orthonormal systems*, "Nauka", Moscow, 1971; English transl., Transl. Math. Monographs, vol. 29, Amer. Math. Soc., Providence, R. I., 1977.
27. —, *Comments on the proof of the conjecture on logarithmic coefficients*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 109–112.
28. R. Nevanlinna, *Erweiterung der Theorie des Hilbertschen Raumes*, Comm. Sémin. Math. Univ. Lund. Tome Supplémentaire (1952), 160–168.
29. —, *Über metrische lineare Räume. II, Bilinearforme und Stetigkeit*, Ann. Acad. Sci. Fenn. Sci. A I Math. **113** (1952).
30. —, *Über metrische lineare Räume. III, Theorie der Orthogonalsysteme*, Ann. Acad. Sci. Fenn. Sci. A I Math. **115** (1952).
31. —, *Über metrische lineare Räume. IV, Zur Theorie der Unterräume*, Ann. Acad. Sci. Fenn. Sci. A I Math. **163** (1954).
32. N. K. Nikol'skii, *A treatise on the shift operator*, "Nauka", Moscow, 1980; English transl., Springer-Verlag, Heidelberg, 1985.
33. N. K. Nikol'skii and V. I. Vasyunin, *Notes on two functional models*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 113–141.
34. M. Ozawa, *An elementary proof of the Bieberbach conjecture for the sixth coefficient*, Kodai Math. Seminar Reports **21** (1969), 129–132.
35. R. N. Pederson, *A proof of the Bieberbach conjecture for the sixth coefficient*, Arch. Rational Mech. Anal. **31** (1968), 331–351.

36. R. Pederson and M. Schiffer, *A proof of the Bieberbach conjecture for the fifth coefficient*, Arch. Rational Mech. Anal. **45** (1972), 161–193.
37. H. Prawitz, *Über Mittelwerte analytischer Funktionen*, Arkiv für Mat. Astron. Fysik **20** (1927–1928), no. 6, 1–12.
38. M. S. Robertson, *A remark on the odd schlicht functions*, Bull. Amer. Math. Soc. **42** (1936), 366–370.
39. M. Rosenblum and J. Rovnyak, *Hardy classes and operator theory*, Clarendon Press, Oxford, 1985.
40. D. Sarason, *Generalized interpolation in H^∞* , Trans. Amer. Math. Soc. **127** (1967), 179–203.
41. —, *Invariant subspaces from the Brangesian point of view*, The Bieberbach Conjecture: Proceedings of the Symposium on the Occasion of the Proof, Math. Surveys Monogr., no. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 153–166.
42. I. Schur, *On Faber polynomials*, Amer. J. Math. **67** (1945), 33–41.
43. —, *Ein Satz über quadratische Formen mit komplexen Koeffizienten*, Amer. J. Math. **67** (1945), 472–480.
44. —, *Identities in the theory of power series*, Amer. J. Math. **69** (1947), 14–26.
45. B. Sz-Nagy and C. Foias, *Analyse: Harmonique des opérateurs de l'espace de Hilbert*, Masson, Paris, 1967.
46. —, *On the structure of intertwining operators*, Acta Sci. Math. (Szeged) **35** (1973), 225–254.
47. O. Tammi, *Extremum problems for bounded univalent functions. I*, Lecture Notes in Math., vol. 646, Springer-Verlag, Heidelberg, 1978.
48. —, *Extremum problems for bounded univalent functions. II*, Lecture Notes in Math., vol. 913, Springer-Verlag, Heidelberg, 1982.

PURDUE UNIVERSITY, LAFAYETTE, INDIANA 47906, USA

The Geometry of 4-Manifolds

SIMON K. DONALDSON

I. Introduction. The title of this lecture is appropriate because, while the results we describe lie in the field of differential topology, the methods used are geometrical, exploiting the “instantons” or “Yang-Mills fields” introduced by physicists. Before going on to a detailed survey of results and techniques we will first contrast these developments with the general pattern of manifold topology.

A topological n -manifold is constructed from domains in n -dimensional Euclidean space, pieced together by homeomorphisms. The manifold is provided with a differentiable or smooth structure if these homeomorphisms are differentiable. The basic equivalence relation among topological manifolds is that of *homeomorphism* and among smooth manifolds is *diffeomorphism* (homeomorphism defined by smooth functions). In the 1960s and 1970s topologists developed a comprehensive theory of manifolds in dimension 5 or more. This theory explained the relationship between the smooth and topological categories [28]; and for many classes of manifolds it gave a complete classification in terms of invariants from algebraic topology [3, 26, 33]. The coarsest of these are *homotopy invariants*, for example, the homology groups. Next most important are the *Pontrayagin classes* $p_i(X)$ in $H^{4i}(X; \mathbf{Z})$ of a smooth manifold X —characteristic classes of the tangent bundle. Let us focus on four facts from this high-dimensional theory:

(1) Simply connected smooth manifolds of dimension 5 or more are diffeomorphic if they are h -cobordant.

(2) The homotopy type and Pontrayagin classes of a compact simply connected manifold of dimension 5 or more determine the smooth structure up to a finite number of possibilities.

(3) The reductions of the Pontrayagin classes to $H^*(X; \mathbf{Q})$ are topological invariants.

(4) A contractible topological manifold of dimension 5 or more has a unique smooth structure.

In high dimensions the h -cobordism technique gives an effective method for constructing equivalences between manifolds, and the classification of smooth and topological manifolds differ by only a “finite amount.”

In 1982 Freedman [16] showed that the basic constructions used in high dimensions could be carried out with *topological* 4-manifolds. The sole classical invariant of a compact simply connected 4-manifold is the intersection form on 2-dimensional homology. By Hirzebruch's theorem the first Pontrayagin number $p_1(X^4)[X^4]$ is 3 times the signature $b_2^+ - b_2^-$, where b_2^+ and b_2^- are the dimensions of the positive and negative parts of this quadratic form. Freedman's theory asserts that, up to a finite ambiguity, the topological classification of 4-manifolds mimics the algebraic classification of forms.

Now among recent results for smooth 4-manifolds we have:

(1)' There are simply connected, smooth 4-manifolds which are h -cobordant but nondiffeomorphic.

(2)' There is a countably infinite family of smooth, simply connected 4-manifolds, all mutually homeomorphic but with distinct smooth structures.

(3)' There are rational cohomology invariants of smooth 4-manifolds which (unlike the Pontrayagin classes) depend essentially on the smooth structure.

(4)' There is an uncountable family of smooth 4-manifolds, each homeomorphic to \mathbf{R}^4 but with mutually distinct smooth structures.

(See §III below for more precise statements and references.) These facts, all coming from Yang-Mills theory, emphasize the very different picture we are beginning to see in four dimensions.

II. Techniques.

(i) *The first order Yang-Mills equations.* These equations in 4-dimensional geometry are in some ways analogous to the Cauchy-Riemann equations in dimension 2. In place of the functions on a Riemann surface the basic geometric objects we take are the *connections* on a bundle E over an oriented Riemannian 4-manifold X . The structure group of E is some compact Lie group G , for example $SU(2)$ or $SO(3)$. In place of the splitting of the derivative of a function into holomorphic and antiholomorphic parts we have the splitting of the curvature F_A of a connection A into self-dual and anti-self-dual parts: $F_A = F_A^+ + F_A^-$. These are the components in the eigenspaces of the Hodge $*$ operator, acting on bundle-valued 2-forms. In place of the holomorphic functions we have the *anti-self-dual connections* (or *instantons*), solutions of the equation $F_A^+ = 0$. Geometrically this condition means that the curvature F_A takes opposite values on any pair of orthogonal 2-planes in the tangent space of X .

This anti-self-dual equation is a first-order partial differential equation for the connection A . It has a large group of symmetries: the group of automorphisms or "gauge transformations" of the bundle E . When this is taken into account (by identifying solutions which differ by a gauge transformation) the equation becomes elliptic. It depends only on the conformal class of the Riemannian metric on X . Many of its special features spring from a fundamental identity linking the "energy" of a solution with the topology of the bundle E . If X is

compact and A is anti-self-dual, then

$$\int_X |F_A|^2 d\mu = c(G) \cdot k(E), \quad (1)$$

where $c(G)$ is a normalizing constant and $k(E)$ is an integer—a characteristic number of E . For example, if G is $\mathrm{SO}(3)$, then k is minus the first Pontrayagin class of E , evaluated on X . Again, analogous identities hold for the energy of holomorphic maps.

(ii) *Nonlinear Fredholm theory.* Differential topology in infinite-dimensional manifolds provides a convenient language to describe many properties of the Yang-Mills instantons, primarily those stemming from the implicit function theorem in Banach spaces. This is applied to nonlinear differential operators between suitable Sobolev spaces. For a fixed bundle $E \rightarrow X$ one defines a space \mathcal{B}_E of all gauge equivalence classes of connections, orbits under the group of gauge transformations of E . An open dense subset \mathcal{B}_E^* of \mathcal{B}_E is an infinite-dimensional manifold; its complement, the singular set of \mathcal{B}_E , represents *reducible connections* with holonomy group a subgroup of G whose centralizer properly contains the center of G .

The second stock of ideas which can be applied are those based on Sard's theorem and transversality. As Smale observed [30] these basic constructions of differential topology carry over to infinite-dimensional problems involving *Fredholm mappings*: smooth maps whose derivatives have finite-dimensional kernels and cokernels. As an illustration (particularly relevant to the definition of the invariants in §III(iv) below) consider a Fredholm map $\varphi: E \rightarrow F$ between Banach spaces whose *index* (the integer $\dim(\ker d\varphi)_x - \dim(\mathrm{coker} d\varphi)_x$, calculated for any x in E) is zero. Generic points y in F are regular values of φ and for these $\varphi^{-1}(y)$ is a discrete subset of X . If φ is a proper map, then this set is finite. We can attach a sign to each point, in such a way that the algebraic sum over the fibers yields an integer invariant, independent of y . The proof in the general case is not significantly different from that in finite dimensions. In the same way this integer—the “degree” of φ —is a deformation invariant unchanged by proper Fredholm homotopies. More generally, if E is replaced by a Banach manifold B we can associate homology classes in $H_d(B; \mathbf{Z})$ to suitable Fredholm maps φ , where d is the index of φ .

The anti-self-dual equations fit into this framework. The *moduli space* M_E (space of equivalence classes of the equation $F_A^+ = 0$) is defined by a Fredholm mapping whose index was calculated by Atiyah, Hitchin, and Singer [2] (applying the Atiyah-Singer index theorem to the linearized operator). They gave a general formula:

$$\dim M_E = 2a_G \cdot k(E) - \dim G(1 - b_1(X) + b_2^+(X)), \quad (2)$$

where a_G is an integer depending only on G (equal to 1 when $G = \mathrm{SO}(3)$, for example). This is the “virtual dimension” of M_E , and typically one expects the part of the moduli space in \mathcal{B}_E^* to be a smooth manifold of this dimension. More precisely, Freed and Uhlenbeck prove in [15] that for nontrivial $\mathrm{SU}(2)$ and $\mathrm{SO}(3)$

bundles E this is the case for typical metrics on X . In general one can achieve the same end by making small perturbation of the anti-self-duality equations. Note here that the parity of $\dim M_E$ is *independent of the bundle E* .

The same kinds of ideas can be applied to the reducible connections, the nonmanifold points in \mathcal{B} . That is, we can describe the behavior of the moduli spaces there in typical situations. The important reductions are those with abelian structure group S^1 , and these can be studied by Hodge theory. As the Riemannian metric on X varies, these reducible solutions to the anti-self-dual equations appear on subsets of codimension b_2^+ . So we can avoid them in families of metrics of dimension less than b_2^+ . In families of dimension b_2^+ we encounter a fundamental singularity in the associated moduli spaces. For example, if X has a negative definite intersection form, so $b_2^+ = 0$, the singularities are always present and, for typical metrics, are cones on complex projective spaces.

(iii) *Compactification*. The Yang-Mills moduli spaces are not, in general, compact, but a theorem of Uhlenbeck singles out a natural compactification. This control “at infinity” in the space \mathcal{B}_E of connections stands in for the properness of the Fredholm map defining the moduli spaces, which holds only in special cases.

Uhlenbeck’s theorem [37] supplies information on connections given bounds on their energy. For anti-self-dual connections these come from the identity (1). Let us restrict for simplicity to $\mathrm{SO}(3)$ bundles E , which are determined topologically by characteristic classes $p_1(E)$ in $H^4(X; \mathbf{Z}) \cong \mathbf{Z}$ and $w_2(E)$ in $H^2(X; \mathbf{Z}/2)$ with $w_2^2 = p_1 \bmod 4$. So if we fix $w_2 = U$ there is a family of moduli spaces, $M_j = M_{j,U}$ say, with $j \geq 0$. Then one can define a topology on

$$M_j \cup X \times M_{j-4} \cup S^2(X) \times M_{j-8} \cup \cdots$$

such that the closure \overline{M}_j of M_j is compact. Here the $S^i(X)$ denote the *symmetric products* of i points in X . The points in the lower “strata” $S^i(X) \times M_{j-4i}$ represent “ideal connections” whose energy density $|F_A|^2$ is augmented by δ -functions at i points in X .

Thanks to work of Taubes [34], extended in [7], we have a good hold on the structure of neighborhoods of the lower strata in \overline{M}_k , that is, of the “ends” of the moduli spaces. By making a detailed analysis of the relevant implicit function theorem one describes neighborhoods of $S^i(X) \times M_{j-4i}$ in M_j in terms of a connection in M_{j-4i} , i copies of the “fundamental instanton” at points of X , and “glueing data” which identifies these component parts. For example, the link of $X \times M_{j-4}$ in \overline{M}_j is typically a copy of the structure group $\mathrm{SO}(3)$.

Ideas of this kind, describing the behavior of differential operators “at infinity” in a function space, have appeared recently in a number of different geometric problems. In gauge theory, Taubes has used them to construct a calculus of variations—see Taubes’s lecture at this Congress.

(iv) *The anti-self-dual equations and holomorphic geometry*. Suppose the base space X is a 2-dimensional complex surface with a Hermitian metric. If E is a complex vector bundle over X (with structure group a subgroup of the

unitary group), any connection defines an almost complex structure on E . If the connection is anti-self-dual its curvature has type $(1, 1)$ and this implies that the structure is integrable. We get, in this way, a map from the anti-self-dual connections over X to the holomorphic bundles, the latter depending only on the complex geometry of X .

A holomorphic bundle is, by definition, locally trivial whereas there are many local solutions to the anti-self-dual equations. However globally we can reconstruct the connection from its holomorphic bundle. There is nothing special here about the dimension of the base space. If (Y, ω) is any compact Kahler manifold and $E \rightarrow Y$ a holomorphic bundle with structure group $\mathrm{SL}(r, \mathbb{C})$ (say), any metric on E determines a reduction of the structure group to $\mathrm{SU}(r)$ and also a preferred $\mathrm{SU}(r)$ connection. We look for metrics such that the curvature F of this connection satisfies

$$F \cdot \omega = 0 \tag{3}$$

at every point of Y . This is a second-order elliptic equation for the metric on E . On the other hand, in algebraic (or holomorphic) geometry there is a notion of a *stable* vector bundle, introduced by algebraic geometers in moduli problems. We have:

PROPOSITION. *The holomorphic bundle E is stable if and only if it carries an irreducible solution of the differential equation (3). The solution is then unique.*

This was proved recently by Uhlenbeck and Yau [38]. The result had been conjectured (independently) by Hitchin and Kobayashi; in the simplest case when Y is a complex curve, it is equivalent to a theorem of Narasimhan and Seshadri, and this was developed from the point of view of Yang-Mills theory by Atiyah and Bott [1]. For an algebraic surface Y the result was proved in [6].

So on compact Kahler manifolds of any dimension, this theorem of Uhlenbeck and Yau gives a holomorphic description of the unitary connections whose curvature is of type $(1, 1)$ and perpendicular to the Kahler form. The special feature of complex surfaces is that these are *precisely* the anti-self-dual connections. Thus for algebraic surfaces the moduli spaces M_E can be described using algebraic geometry. They are quasi-projective complex varieties. From this point of view the best algebraic construction is that of Gieseker [20].

III. Results and applications.

(i) *Realizing intersection forms.* Here we discuss results forbidding the construction of smooth 4-manifolds with given intersection forms. They can be seen alternatively as obstructions to smoothing the topological manifolds constructed by Freedman. Equally, they imply that it is impossible to do smooth *surgery* on homology classes in many existing manifolds.

The first theorem of this kind asserted that nonstandard (nondiagonalizable) definite forms cannot be realized by smooth, simply connected 4-manifolds [5]. The proof used a 5-dimensional moduli space of $\mathrm{SU}(2)$ connections. The result has since been extended in two different ways.

On the one hand, Fintushel and Stern found a rather simple proof for negative definite forms which represent -2 or -3 . They considered $\mathrm{SO}(3)$ connections, choosing w_2 and p_1 to get moduli spaces of low dimensions [12]. Their proof dealt with manifolds with no 2-torsion in H_1 . In both proofs the moduli spaces are truncated to give manifolds with a known boundary. Boundary contributions come from the links of singularities at reducible connections and from the lower strata in the compactified moduli space. Then one asserts that the moduli manifold gives a cobordism, and a fortiori homology, in \mathcal{B}_E^* between these boundaries.

On the other hand, the proof of [5] was extended by Furuta [18] and the author [8] to take account of fundamental group. This required a more extensive use of transversality and also a detailed study of the *orientation* of the moduli spaces. The upshot is the optimal result for definite forms:

THEOREM 1 [8]. *If a smooth, compact, oriented 4-manifold has a definite intersection form, then the form can be diagonalized over the integers.*

There are also results for some indefinite forms [7]. These are proved in a similar way, using more complicated analysis and topology. One can define a map

$$\mu: H_2(X; \mathbf{Z}) \rightarrow H^2(\mathcal{B}_E^*; \mathbf{Z}) \quad (4)$$

by decomposing the 4-dimensional characteristic class of the “universal” bundle over $\mathcal{B}_E^* \times X$. For indefinite manifolds the moduli spaces typically avoid the reductions, so, by restricting μ , we construct cohomology classes over the moduli spaces. For the same reason the only boundary contributions are now those from the lower strata. There are further, mod 2, cohomology classes which detect the links of the lower strata in the homology of \mathcal{B}_E^* . The best result so far is

THEOREM 2 [7]. *If a smooth, compact, oriented 4-manifold has no 2-torsion in H_1 and an even intersection form with a positive part of rank 2, then the form is*

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The method appears to run out of steam as b_2^+ grows because the relative size of the contributions to the ends from different lower strata changes.

(ii) *Orbifolds and the representation of homology classes.* Fintushel and Stern began the study of Yang-Mills equations on 4-dimensional orbifolds: spaces with a discrete set of singularities modelled on finite quotients of \mathbf{R}^4 . These occur naturally as the quotients of smooth 4-manifolds by finite groups or of 5-manifolds by circle actions. They are rational homology manifolds, and analysis on them is quite similar to that on smooth manifolds—the chief modification is the appearance of extra terms in the index formula (2) owing to the singularities. Using variants of their argument for manifolds, Fintushel and Stern obtained restrictions on the existence of orbifolds with certain intersection forms and singularities. Their results have many applications, notably to the group θ_H^3 of

homology 3-spheres modulo homology cobordism. Before Fintushel and Stern's work it had seemed possible that this group was rather small, perhaps of order 2. In fact we have

THEOREM 3 [13]. *The Poincaré homology sphere P has infinite order in θ_H^3 (i.e., no connected sum $P \# \cdots \# P$ bounds an acyclic smooth 4-manifold). Moreover $\theta_H^3 / \langle P \rangle$ is nonzero.*

(By contrast, the Poincaré sphere itself bounds an acyclic topological 4-manifold [16].)

Another application was to give an alternative proof of a theorem of Kuga. Any 2-dimensional homology class in a 4-manifold can be represented by a smoothly embedded surface. It is an interesting general problem to find lower bounds on the genus of such a representative. Kuga's theorem considers classes in $H_2(S^2 \times S^2)$, written in the standard basis as pairs (p, q) of integers.

THEOREM 4 [23]. *The class (p, q) in $H_2(S^2 \times S^2)$ can be represented by a smoothly embedded 2-sphere if and only if either p or q is 0, +1, or -1.*

(By contrast, if p and q are co-prime the class can be represented by a topologically flat embedded sphere.)

Kuga's original proof was indirect, applying the results of §III(i).

(Similar arguments for other manifolds have been made by Lawson [25] and Suciu [31].) Fintushel and Stern gave a simpler proof using orbifolds. If a 2-sphere, embedded in a 4-manifold with nonzero self-intersection number, is collapsed to a point, the resulting space is an orbifold (whose singularity is a cone on a lens space). Later Furuta [19] gave an even more direct proof using other moduli spaces on these orbifolds. In another direction Lawson [24] used these techniques to study embedded projective planes.

The techniques in Fintushel and Stern's first paper have recently been extended by Fintushel, Lawson, and Stern. One application yields results on the exceptional orbits of circle actions on S^5 , partially proving a conjecture of Montgomery and Yang [14].

(iii) *Exotic structures on \mathbf{R}^4 .* Freedman's theory asserts that direct sum decompositions of the intersection form of a 4-manifold can be realized, by surgery, as topological decompositions of the manifold. The results of §III(i) prevent these being made smoothly. This conflict implies that there exist "exotic \mathbf{R}^4 's": smooth manifolds homeomorphic but not diffeomorphic to Euclidean space [21]. The first examples were open subsets of $S^2 \times S^2$ or \mathbf{CP}^2 . The proofs of their exotic nature were indirect. Later Gompf found a countably infinite family of exotic \mathbf{R}^4 's in this way.

Dramatic further progress was made by Taubes [35], carrying out a program suggested by Freedman. Taubes extended the fundamentals of Yang-Mills theory to "end-periodic" 4-manifolds. These are noncompact manifolds whose end has a periodic configuration $W_1 \cup W_2 \cup \cdots \cup W_n \cup \cdots$, where the W_i are overlapping

copies of an open manifold W . The simplest examples are manifolds whose end is a tube $Y^3 \times (0, \infty)$ and W is $Y \times (0, 1)$. Taubes showed that given conditions

(1) on the homology of W and

(2) on the *representations* $\pi_1(W) \rightarrow \mathrm{SU}(2)$,

these behave like compact manifolds from the point of view of the anti-self-dual equations. By substituting the resulting theorems on the intersection forms of end-periodic manifolds into Freedman's analysis of the failure of smooth surgery Taubes proved:

THEOREM 5 [22, 35]. *There exists a family $R_{s,t}$ of smooth 4-manifolds, parametrized by $(s, t) \in \mathbf{R}^2$, each homeomorphic to \mathbf{R}^4 but no two diffeomorphic.*

Thus there are "moduli" of smooth structures on the topological manifold \mathbf{R}^4 . Moreover Taubes's family does not contain all exotic \mathbf{R}^4 's. None of the $R_{s,t}$ can be embedded in the standard \mathbf{R}^4 , but the failure of the h -cobordism theorem (§III(iv)) implies that examples with this property do exist.

In a similar spirit to this work of Taubes, the author and Sullivan have extended the fundamentals of Yang-Mills theory to *quasi-conformal* 4-manifolds, whose local co-ordinates compare by quasi-conformal maps of domains in \mathbf{R}^4 [11]. The work is allied to that of Teleman on Lipschitz manifolds [36]. Essentially all the results proved for smooth 4-manifolds and diffeomorphisms, using the anti-self-duality equations, extend to quasi-conformal manifolds and quasi-conformal maps. (In particular there are exotic quasi-conformal structures on \mathbf{R}^4 .) A fortiori the results extend to Lipschitz 4-manifolds. This is in sharp contrast with the theorem of Sullivan [32]: in high dimensions every topological 4-manifold has a unique Lipschitz structure.

(iv) *New invariants.* The results here are the other side of the coin displayed in §III(i). Many compact topological 4-manifolds cannot be smoothed: those that can may carry many different smooth structures. This is established by constructing differential topological invariants from the Yang-Mills moduli spaces, along the lines indicated in §II(i).

Let us restrict attention to simply connected 4-manifolds. For any bundle E over X the rational cohomology of \mathcal{B}_E^* is generated as a ring by classes c/α , where c is a rational characteristic class of the universal bundle on the product $\mathcal{B}_E^* \times X$ and α is a homology class in X . In particular, all the rational cohomology of \mathcal{B}_E^* lies in *even dimensions*. For example, if G is $\mathrm{SO}(3)$, then $H^*(\mathcal{B}_E^*; \mathbf{Q})$ is a polynomial algebra, generated by the image of the map μ in (4) and a further class in H^4 . Since the invariants we expect to see with the ideas of §II(ii) lie, roughly speaking, in the *homology* of \mathcal{B}_E^* we anticipate good results in cases when the moduli spaces are *even-dimensional*, and by (2) this happens exactly when $b_2^+(X)$ is *odd*.

Manifolds with $b_2^+ = 1$ form a rather special class here, since reducible solutions appear in the moduli space for a codimension 1 family of metrics. We consider the two-dimensional moduli space of $\mathrm{SU}(2)$ connections with Chern class 1 ($\mathrm{SO}(3)$ connections with Pontrayagin class -4) over such a manifold X . This

will not always be compact, but there is a way to introduce a correction term for the boundary. Then one can associate to generic metrics on X a homology class in \mathcal{B}_E^* which changes only through the appearance of reducible solutions. In the end one obtains a differential topological invariant of X having the form of a map:

$$\Gamma_X: \mathcal{C}_X \rightarrow H^2(X; \mathbf{Z}),$$

where \mathcal{C}_X is a set of “chambers” in $H^2(X; \mathbf{R})$ [9].

For algebraic surfaces X one can hope to calculate this invariant using the holomorphic description of anti-self-dual connections via stable bundles. Any rational surface has $b_2^+ = 1$, but there are also irrational examples, in particular, a family $D_{p,q}$ (p, q coprime integers) constructed by Dolgachev. The difference in the complex geometry of the rational and irrational manifolds is reflected in the stable bundles and so in the moduli spaces and the invariant Γ . This gives

THEOREM 6 [9]. *A Dolgachev surface $D_{p,q}$ is homotopy equivalent (hence homeomorphic and h -cobordant) but not diffeomorphic to a connected sum $\mathbf{CP}^2 \# 9\mathbf{CP}^2$.*

So the h -cobordism theorem does not extend to smooth 4-manifolds. Going further, Friedman and Morgan and Okonek and Van de Ven used the Γ -invariant to prove:

THEOREM 7 [17, 27]. *There are infinitely many diffeomorphism types among the homotopy equivalent manifolds $D_{p,q}$.*

When b_2^+ is odd and bigger than 1, many other invariants can be defined. For any bundle E the moduli space is of even dimension $2d(E)$. The classes $\mu(\alpha)$, for α in $H_2(X)$, can be represented by cochains with “small” support in \mathcal{B}_E^* . Our description of the end of the moduli space then allows the construction (in a stable range $k(E) \gg 0$) of a pairing between the powers $\mu(\alpha)^d$ and the fundamental class of M_E . Considering $\mathrm{SO}(3)$ connections one gets:

THEOREM 8 [10]. *Let X be a simply connected, smooth, oriented 4-manifold with $b_2^+(X) = 2p + 1$, $p > 0$. Fix an orientation of a maximal positive subspace for the intersection form on H^2 . Then for any u in $H^2(X; \mathbf{Z}/2)$ with $u^2 = \alpha \bmod 4$ and for $j > j_0(p)$ the homology class of the $\mathrm{SO}(3)$ moduli space $M_{u,j}$ defines a polynomial*

$$q_{u,j,X}: S^d(H_2(X)) \rightarrow \mathbf{Z}$$

of degree $d = j - 3(1 + p)$, independent of the metric on X .

(Here the orientation of the positive subspace orients the moduli space M_E .) So for roughly “half” of the possible simply connected 4-manifolds we can define *infinitely many* new invariants. At present they are very hard to calculate; their main application has come from the tension between two general properties of

the moduli spaces. On the one hand, we have for connected sums a “vanishing theorem”:

THEOREM 9 [10]. *If the 4-manifold X of Theorem 8 is a connected sum $X_1 \# X_2$ with each $b_2^+(X_i) > 0$, then all the invariants $q_{u,j,X}$ are 0.*

On the other hand, if X is a projective algebraic surface there is a preferred “hyperplane” class $[H]$ in $H_2(X)$. For large values of j the moduli spaces are quasi-projective varieties of the “proper” dimension. Gieseker’s construction shows that $\mu(H)$ is the first Chern class of an ample line bundle over the moduli space. So $\langle \mu(H)^d, M \rangle$ is positive and we deduce

THEOREM 10 [10]. *If a simply connected compact complex algebraic surface can be written as a connected sum, then the intersection form of one of the summands is negative definite.*

This immediately gives many more examples of manifolds, with the same classical invariants, distinguished by the new invariants $q_{u,j,X}$.

IV. Problems. The techniques described here are a long way from becoming a systematic theory. One notable feature is that, while the only known proofs of the ten theorems in §III use Yang-Mills instantons, there are, in most cases, a number of alternative proofs available (arguing with different moduli spaces, etc.). This suggests that there may be some more fundamental principle, relating 4-manifold topology with Yang-Mills theory, of which these arguments are different manifestations. If we could find such a principle, it might point the way to attack problems which seem to lie beyond the methods discussed above. The most obvious general questions are:

(1) Which even indefinite forms are the intersection forms of smooth, simply connected 4-manifolds? (The simplest open case is the rank 38 form $4E_8 \oplus 3\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.)

(2) In which homotopy types are there compact, simply connected, 4-manifolds with distinct smooth structures?

For (oriented) manifolds with b_2^+ odd there are many new invariants with which one can hope to distinguish smooth structures, so we are led to ask:

(3) Are there homotopy equivalent, simply connected, 4-manifolds with b_2^+ even having distinct smooth structures? The smooth 4-dimensional Poincaré conjecture is an instance of this.

On the other hand many problems to do with our new invariants for manifolds with b_2^+ odd present themselves:

(4) Are there universal relations among the invariants $q_{k,u,X}$?

(5) Can we systematically calculate the invariants given some standard description of a 4-manifold?

Two avenues seem to be promising. First we have the holomorphic description of the moduli spaces when the base manifold is a complex surface. Perhaps there are general relations between our invariants and the usual algebro-geometric

invariants of surfaces. A concrete question is:

(6) If X is a minimal algebraic surface, are the invariants $q_{k,u,X}$ of Theorem 8 all polynomials in the canonical class $c_1(K_X)$ and the intersection form of X ?

An extreme possibility is that they are given by universal polynomials in these two variables, with coefficients depending on k .

Second, a new slant on the picture in 4 dimensions may come from the work of Casson [4]. He defines an integer invariant for homology 3-spheres Y^3 using the representations $\pi_1(Y^3) \rightarrow \mathrm{SU}(2)$. The Casson invariant can be calculated from a Dehn surgery description of Y . Now these representations also come to the fore in Taubes's extension of Yang-Mills theory to noncompact (end periodic) 4-manifolds. Recently, Taubes has shown that Casson's invariant can be put into the same framework of Fredholm maps over Banach manifolds described in §II (ii). Moreover Casson and Taubes found different proofs, as corollaries of their work, of

THEOREM 11 [4, 35]. *There exist topological 4-manifolds which are not homeomorphic to a simplicial complex.*

These two proofs are circumstantial evidence for the existence of some link between Casson's invariant and the anti-self-dual equations over 4-manifolds. Perhaps there is a path through Casson's work which will allow our new invariants to be defined using more familiar methods of geometric topology.

ACKNOWLEDGMENT. The author is very grateful to the Universidad del Valle, Cali, Columbia for their hospitality during the preparation of this article.

REFERENCES

1. M. F. Atiyah and R. Bott, *The Yang-Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. London Ser. A **308** (1982), 523–615.
2. M. F. Atiyah, N. J. Hitchin, and I. M. Singer, *Self duality in four dimensional Riemannian geometry*, Proc. Roy. Soc. London Ser. A **362** (1978), 425–461.
3. W. Browder, *Surgery on simply connected manifolds*, Springer-Verlag, Berlin, 1971.
4. A. Casson (to appear).
5. S. K. Donaldson, *An application of gauge theory to 4-dimensional topology*, J. Differential Geom. **18** (1983).
6. —, *Anti-self-dual Yang-Mills connections on complex algebraic surfaces and stable vector bundles*, Proc. London Math. Soc. (3) **50** (1985), 1–26.
7. —, *Connections, cohomology and the intersection forms of four manifolds*, J. Differential Geom. **24** (1986), 275–341.
8. —, *The orientation of Yang-Mills moduli spaces and four-manifold topology*, J. Differential Geom. (to appear).
9. —, *Irrationality and the h-cobordism conjecture*, J. Differential Geom. **25** (1987).
10. —, *Polynomial invariants for smooth 4-manifolds* (in preparation).
11. S. K. Donaldson and D. P. Sullivan (in preparation).
12. R. Fintushel and R. Stern, *$\mathrm{SO}(3)$ connections and the topology of 4-manifolds*, J. Differential Geom. **20** (1984), 523–539.
13. —, *Pseudo-free orbifolds*, Ann. of Math. (2) **122** (1985), 325–364.
14. —, *$O(2)$ actions on the 5-sphere*, Invent. Math. (to appear).
15. D. S. Freed and K. K. Uhlenbeck, *Instantons and 4-manifolds*, Math. Sci. Res. Inst. Pub., Springer-Verlag, New York, 1984.

16. M. H. Freedman, *The topology of 4-dimensional manifolds*, J. Differential Geom. **17** (1982), 357–453.
17. R. Friedman and J. W. Morgan, *On the diffeomorphism types of certain algebraic surfaces*, submitted to J. Differential Geom.
18. M. Furuta, *Perturbation of moduli spaces of self-dual connections*, submitted to J. Fac. Sci. Univ. Tokyo Sect. I A Math.
19. —, *On self-dual pseudo-connections on some orbifolds*, Preprint, Univ. of Tokyo.
20. D. Gieseker, *On the moduli of vector bundles over an algebraic surface*, Ann. of Math. (2) **106** (1977), 45–60.
21. R. Gompf, *Three exotic \mathbf{R}^4 's and other anomalies*, J. Differential Geom. **18** (1983), 317–328.
22. —, *An infinite set of exotic \mathbf{R}^4 's*, J. Differential Geom. **21** (1985), 283–300.
23. K. Kuga, *Representing homology classes in $S^2 \times S^2$* , Topology **23** (1984), 133–137.
24. T. Lawson, *Normal bundles for an embedded \mathbf{RP}^2 in a positive definite 4-manifold*, J. Differential Geom. **22** (1985), 215–231.
25. —, *Representing homology classes in an almost definite 4-manifold*, Preprint.
26. S. P. Novikov, *Homotopy equivalent smooth manifolds*, Izv. Akad. Nauk. SSSR. Ser. Mat. **28** (1964), 365–474; English transl. in Amer. Math. Soc. Transl. (2) **48** (1965).
27. C. Okonek and A. Van de Ven, *Stable bundles and differentiable structures on certain elliptic surfaces*, Invent. Math. **86** (1986), 357–370.
28. L. C. Siebenmann, *Structures on topological manifolds*, Proc. Internat. Congr. Math. (Nice, 1970), Vol. 2, Gauthier-Villars, Paris, 1971, pp. 133–163.
29. S. Smale, *On the structure of manifolds*, Amer. J. Math. **84** (1964), 387–399.
30. —, *An infinite dimensional version of Sard's Theorem*, Amer. J. Math. **87** (1965), 861–866.
31. A. I. Suciu, *Immersed spheres in \mathbf{CP}^2 and $S^2 \times S^2$* , Preprint, Yale University.
32. D. P. Sullivan, *Hyperbolic geometry and homeomorphisms*, Geometry Topology (Proc. Georgia Topology Conf., Athens, Ga., 1977), J. Cantrell, editor, Academic Press, New York, 1978, pp. 543–555.
33. —, *Infinitesimal computations in topology*, Inst. Hautes Études Sci. Publ. Math. **47** (1977), 269–332.
34. C. H. Taubes, *On the existence of self-dual connections on manifolds with indefinite intersection matrix*, J. Differential Geom. **19** (1984), 517–560.
35. —, *Gauge theory on asymptotically periodic 4-manifolds*, Preprint, Harvard Univ., Cambridge, Mass.
36. N. Teleman, *The index of signature operators on Lipschitz manifolds*, Inst. Hautes Études Sci. Publ. Math. **58** (1983), 39–79.
37. K. K. Uhlenbeck, *Connections with L^p -bounds on curvature*, Comm. Math. Phys. **83** (1982), 11–30.
38. K. K. Uhlenbeck and S. T. Yau, *On the existence of Hermitian Yang-Mills connections on stable bundles*, Preprint, University of Chicago.

THE MATHEMATICAL INSTITUTE, OXFORD, ENGLAND

Neuere Entwicklungen in der arithmetischen algebraischen Geometrie

GERD FALTINGS

Ich möchte einige Höhepunkte in der Entwicklung der arithmetischen algebraischen Geometrie in den letzten vier Jahren darstellen. Naturgemäß ist die Auswahl rein subjektiv, und die Darstellung erhebt keinen Anspruch auf Vollständigkeit. Wir konzentrieren uns auf die folgenden vier Themen:

- (a) Die Formel von Gross-Zagier,
- (b) Die Vermutungen von Tate, Schafarevitsch, und Mordell,
- (c) Hodge-Tate Strukturen auf der p -adischen Kohomologie,
- (d) Die Beilinson-Vermutung.

Zu zweien von den obigen Themenkreisen habe ich persönlich Beiträge geleistet, und ich darf daher vielleicht mit einigem Recht versuchen, die Ergebnisse vor diesem großen Kreis von Zuhörern darzustellen. Bei den beiden anderen Gebieten muß mich darauf beschränken, das wiederzugeben, was ich in letzter Zeit aus verschiedenen Quellen gelernt habe. Meine Entschuldigung für diese sicherlich unvollkommene Darstellung besteht darin, daß die Plenar-Vorträge doch für einen weiteren Zuhörer-Kreis als die Sektions-Vorträge bestimmt sind. Für eine kompetentere Darstellung kann ich aber nur auf diese verweisen.

1. Die Formel von Gross-Zagier. Als Hintergrund dienen zwei Schwierigkeiten, die oft in der diophantischen Geometrie auftreten. Die eine ist die Bestimmung der Nullstellen-Ordnung von L -Reihen. Dies sind Verallgemeinerungen der klassischen L -Reihen. Es handelt sich um Dirichlet-Reihen $\sum a_n \cdot n^{-s}$, welche in einer Halbebene $\operatorname{Re}(s) \geq s_0$ konvergieren. Die L -Reihe gehört zu einer arithmetischen Varietät X , und die Koeffizienten a_n berechnen sich aus dem Reduktions-Verhalten von X in den endlichen Stellen. Es wird vermutet, daß die L -Reihen sich zu meromorphen Funktionen auf der ganzen komplexen Ebene fortsetzen, und sogar eine Funktional-Gleichung erfüllen. Dies ist gezeigt worden für eine Reihe von arithmetischen Varietäten X , vor allem für gewisse Shimura-Varietäten. Weiter gibt es Vermutungen, daß die Ordnungen der Nullstellen

beziehungsweise Pole der L -Reihe an ganzzahligen Stellen mit geometrischen Invarianten zusammenhängen. Es handelt sich um die Vermutung von Birch und Swinnerton-Dyer und allgemeiner um die Tate-Vermutung.

Da numerische Berechnungen stets mit einem Fehler behaftet sind, kann man mit ihnen nie direkt zeigen, daß eine analytische Funktion an einer bestimmten Stelle verschwindet, sondern höchstens das Gegenteil. Dies macht es schwierig, die oben angeführten Vermutungen zu testen. Dies andere Schwierigkeit betrifft rationale Punkte auf abelschen Varietäten, speziell auf elliptischen Kurven. Man beweist relativ einfach, daß es sich um Torsionspunkte handelt, doch das Gegenteil ist etwas schwieriger. Dabei möchte man oft gerne zeigen, daß die Mordell-Weil Gruppe positiven Rang besitzt, zum Beispiel falls dies von der Vermutung von Birch und Swinnerton-Dyer vorausgesagt wird. Die Formel von B. Gross und D. Zagier erlaubt es nun in einigen Fällen, beide Probleme zu lösen. Sie liefert nämlich die Gleichheit zwischen der Néron-Tate Höhe gewisser rationaler Punkte auf speziellen elliptischen Kurven (Heegner-Punkte auf Weil-Kurven), und der Ableitung einer zugehörigen L -Reihe an der Stelle $s = 1$. Dies kann in beiden Richtungen ausgenutzt werden: Ist der rationale Punkt ein Torsionspunkt (leicht zu zeigen), so verschwindet die Ableitung der L -Reihe (schwierig). Umgekehrt: Ist die Ableitung der L -Reihe verschieden von Null (leicht), so haben wir einen Punkt unendlicher Ordnung gefunden (schwierig).

Die bekannteste Anwendung bis jetzt ist die Bestimmung effektiver unterer Schranken für die Klassenzahlen imaginär quadratischer Zahlkörper. Dies folgt aus einer älteren Arbeit von D. Goldfeld, vorausgesetzt man findet eine L -Reihe wie oben mit einer Nullstelle der Ordnung mindestens drei. Dieses aber leistet gerade die Formel von Gross und Zagier.

Alles in allem handelt es sich um eine schöne Entdeckung, welche wir aber leider noch nicht "erklären" können: Warum ist sie richtig?

Literatur: [G1, G2, O, Zg].

2. Die Mordell-Vermutung. Kommen wir zum Themenkreis der Mordell-Vermutung. Diese besagt, daß auf einer Kurve vom Geschlecht größer als eins über einem Zahlkörper nur endlich viele rationale Punkte liegen. Dies wurde 1922 von L. Mordell vermutet, in der Arbeit in welcher er bewies, daß die rationalen Punkte auf einer elliptischen Kurve ein endlich erzeugte abelsche Gruppe bilden. Allerdings schreibt Mordell selbst, daß er keinen Beweisansatz kennt, und nicht einmal plausibel machen kann warum dies so sein sollte. Die ersten Fortschritte erzielten A. Weil and C. L. Siegel: Der erste verallgemeinerte Mordell's Satz auf abelsche Varietäten, während Siegel mit Hilfe der diophantischen Approximation und A. Weil's Satz die Vermutung für ganze Punkte zeigen konnte. Allerdings führten diese Ansätze im Weiteren nicht mehr zu so großen Fortschritten, und interessanterweise benutzt der endgültige Beweis ganz andere Methoden.

Diese wurden entwickelt zum Beweis der Mordell-Vermutung über Funktionenkörpern. Dies gelang zuerst J. Manin und H. Grauert, und die für uns

wichtige Methode geht zurück auf A. N. Parshin, der die Behauptung reduzierte auf die Schafarevitsch-Vermutung: Über einem Zahlkörper gibt es bis auf Isomorphie nur endlich viele abelsche Varietäten vorgegebener Dimension, welche gute Reduktion außerhalb einer festen endlichen Menge von Stellen des Körpers haben. Die Verbindung zwischen beiden Vermutungen geschieht, indem man zu einem rationalen Punkt auf einer Kurve zunächst eine Überlagerung konstruiert, welche genau über diesem Punkt verzweigt. Auf diese Weise erhält man eine neue Kurve, die schlechte Reduktion höchstens an den Stellen haben kann, wo dies der Fall ist für die ursprüngliche Kurve, plus eventuell einige weitere Stellen (zum Beispiel alle Stellen der Charakteristik zwei). Die Isomorphieklasse dieser Überlagerung bestimmt den rationalen Punkt, und man benötigt nur noch den Satz von Torelli sowie die Vermutung von Schafarevitsch für die Jacobi-Varietät der Überlagerungskurve. Diese Methode liefert übrigens auch den Satz von Siegel über ganze Punkte auf affinen elliptischen Kurven.

Wie aber zeigt man nun die Vermutung von Schafarevitsch? Sie erschien zunächst recht unwahrscheinlich, da über Funktionenkörpern wohl die Mordell-Vermutung gilt, aber nicht diese stärkere Aussage. Trotzdem kommt man bei Zahlkörpern zum Ziel, da man die Tate-Vermutung benutzen kann: Diese reduziert das Problem auf eine Frage über l -adische Darstellungen, und diese kann man einfacher lösen als das ursprüngliche Problem.

Für alles dies benötigt man etwas Information über den Modulraum der prinzipal polarisierten abelschen Varietäten, insbesondere über seine Kompaktifizierungen. Genauer gesagt geht es um die Bestimmung der Höhe einer abelschen Varietät, das heißt des zugehörigen rationalen Punktes im Modulraum. Dazu benötigt man eine Kompaktifizierung des Modulraums, ein amples Geradenbündel darauf, sowie eine hermitesche Metrik für das Geradenbündel. Dies erforderte zunächst einige Tricks, doch mittlerweile besitzen wir eine befriedigende Theorie: Zunächst konstruiert man die toroidale Kompaktifizierung über den ganzen Zahlen. Über dieser dehnt sich die universelle abelsche Varietät aus zu einer semiabelschen Varietät, und man erhält ein Geradenbündel ω , das Determinantenbündel der relativen Differentiale dieser semiabelschen Varietät. Die globalen Schnitte der Potenzen von ω sind gerade die Siegel'schen Modulformen mit ganzzahligen Fourierkoeffizienten. Mit Hilfe von Theta-Reihen ergibt sich, daß eine Potenz von ω global erzeugt ist, und das Bild der korrespondierenden Abbildung von der toroidalen Kompaktifizierung in den projektiven Raum ist die arithmetische Version der minimalen (oder Satake-, oder Baily-Borel-) Kompaktifizierung. Nach Konstruktion gibt es darauf ein kanonisches amples Geradenbündel, und es bleibt die Definition einer guten Metrik. Wieder gibt es nur eine kanonische Wahl, nämlich Quadratintegration von Differentialen. Leider hat diese Metrik Singularitäten im Unendlichen, doch sind diese so mild, daß sie den Gang der Dinge nicht behindern.

Auf diese Weise erhält man eine einfache Definition der Höhe einer abelschen Varietät, mit der sich arbeiten läßt. Es handelt sich im Wesentlichen um das Quadrat-Integral eines Erzeugenden der ganzzahligen Differentiale. Dies liefert

eine numerische Invariante der abelschen Varietät, so daß es bis auf Isomorphie nur endlich viele abelsche Varietäten gibt, für die diese Invariante nicht größer als ein fester vorgegebener Wert ist. Erwähnen wir noch daß die arithmetische Kompaktifizierung des Modulraums auch andere Anwendungen besitzt. Insbesondere ist sie sehr nützlich für eine arithmetische Theorie der Siegelschen Modulformen.

Bleibt schließlich der Beweis der Tate-Vermutung. Genauer gesagt handelt es sich nur um einen Spezialfall der allgemeinen Vermutung, betreffend Endomorphismen abelscher Varietäten. Der Beweis geht zurück auf J. Tate und J. G. Zarhin, und man braucht als neue Ingredienz nur noch die Kompaktifizierung des Modulraums der abelschen Varietäten sowie einige Resultate von J. Tate und M. Raynaud über p -divisible Gruppen. Genauer gesagt betrachtet man eine l -divisible Untergruppe der abelschen Varietät, und dividiert sukzessiv durch die verschiedenen Stufen dieser Untergruppe. Dies liefert eine Folge von abelschen Varietäten, und man muß zeigen, daß unendlich viele dieser isomorph sind. Dazu betrachtet man die zugehörige numerische Invariante, wie weiter oben definiert. Es reicht wenn die Folge dieser reellen Zahlen stationär wird. Es geht also darum, wie sich die Höhe einer abelschen Varietät (so wie oben definiert) bei Isogenien ändert. Man zeigt daß die Variation der Höhe in einer Isogenieklasse beschränkt ist, sogar mit einer ganz effektiv berechenbaren Schranke.

Alles in allem zeigen diese Resultate aufs neue die ungeheure Kraft der Grothendieck'schen Methoden in der algebraischen Geometrie. Man hat aber das Gefühl daß diese Sätze gewissermaßen den vorläufigen Abschluß einer Entwicklung darstellen, und daß wir neue Methoden benötigen, etwa um rationale Punkte auf Varietäten höherer Dimension zu studieren. Dieses wird sicher einiges Experimentieren erfordern, genau wie ehemals die Verhältnisse bei Kurven erarbeitet werden mußten.

Literatur: [D, F2, F3, F4, FW, S1, S2, Z1, Z2].

3. p -adische Darstellungen. Es handelt sich um das Studium der p -adischen étalen Kohomologie einer algebraischen Mannigfaltigkeit über einem p -adischen Körper. Man findet Beziehungen zur kohärenten Kohomologie, und erhält so eine Art von p -adischer Hodge-Theorie. Außerdem gibt es Zusammenhänge mit der kristallinen Kohomologie. Es geht dabei um A. Grothendieck's "mysterious functor."

Versuchen wir zu erklären, worum es sich handelt. Sei K ein p -adischer Körper, etwa eine endliche Erweiterung der p -adischen Zahlen \mathbb{Q}_p . Mit \mathbb{C}_p bezeichnen wir die Kompletterung des algebraischen Abschlusses \bar{K} von K , Darauf operiert stetig die Galois-Gruppe $\mathcal{G} = \text{Gal}(\bar{K}/K)$, und $\mathbb{C}_p(n)$ bezeichnet den Twist mit der n -ten Potenz des zyklotomischen Charakters $\mathcal{G} \rightarrow \mathbb{Z}_p^*$. Als erstes bemerken wir, daß der ungeheuer große Körper \mathbb{C}_p einer Behandlung zugänglich ist: Er enthält den Zwischenkörper, der durch die Einheitswurzeln von p -Potenzordnung erzeugt wird, und ist fast unverzweigt über diesem. Deshalb kann man viele Probleme auf den kleineren Körper zurückführen, wo man sie durch direkte Rechnung löst. Tate und Raynaud haben gezeigt, daß für eine

eigentliche glatte K -Varietät X gilt:

$$H^1(X, \mathbb{Z}_p) \otimes \mathbb{C}_p \cong H^1(X, \mathcal{O}_X) \otimes \mathbb{C}_p \oplus H^0(X, \mathcal{O}_X) \otimes \mathbb{C}_p(-1) \quad (\text{als } \mathcal{G}\text{-Moduln}).$$

Dazu sei bemerkt daß die verschiedenen Twists $\mathbb{C}_p(m)$ nicht isomorph sind, und daß ein \mathbb{C}_p -Vektorraum mit semilinearer \mathcal{G} -Aktion einen größten Unterraum besitzt, welcher isomorph zu einer direkte Summe von $\mathbb{C}_p(m)$'s ist.

Es wurde auch vermutet, und kann inzwischen bewiesen werden, daß allgemeiner gilt:

$$H^n(X, \mathbb{Z}_p) \otimes \mathbb{C}_p \cong \sum_{a+b=n} H^a(X, \Omega_X^b) \otimes \mathbb{C}_p(-b).$$

Insbesondere kann man damit einen rein algebraischen Beweis für das Degenerieren der Hodge Spektral-Sequenz geben. Es folgt auch die Symmetrie der Hodge-Zahlen, zumindest für projektive Varietäten. Bekanntlich reicht dies schon, um viele Resultate zu beweisen, für die man üblicherweise analytische Methoden benutzt. Ein Beispiel ist des Verschwindungssatz von Kodaira.

Die Beweismethode benutzt eine Art von intermediärer Kohomologie, in die sich die Kohomologien auf beiden Seiten der obigen Abbildung natürlich abbilden. Dies reicht, da beide Seiten alle üblichen Axiome erfüllen, wie etwa Künneth-Formel oder Poincaré-Dualität. Die Details sind etwas kompliziert, doch im Wesentlichen kann man die intermediäre Theorie wie folgt beschreiben:

Sei R ein affiner Ring der Varietät, \bar{R} die p -adische Kompletierung der maximalen unverzweigten Erweiterung von R . Dann betrachte man die Galois-Kohomologie von \bar{R} . Die Galois-Kohomologie von \mathbb{Z}_p bildet sich da hinein ab, und dies ergibt die eine natürliche Transformation. Die andere erhält man durch Betrachtung der Differentiale. Einige Details: Wir nehmen an daß R genügend viele Einheiten besitzt. Durch Adjungieren der p -Potenzwurzeln dieser Einheiten erhalten wir eine gut zu kontrollierende Erweiterung \tilde{R} von R , welche unverzweigt ist in Charakteristik 0. Über \tilde{R} ist \bar{R} fast unverzweigt in Kodimension ≤ 2 , und man kann mit einiger Mühe Resultate von \tilde{R} auf \bar{R} übertragen. Sei zum Beispiel \bar{V} die Kompletierung des ganzen Abschlusses des diskreten Bewertungsringes V von K . Die exakte \mathcal{G} -lineare Sequenz

$$0 \rightarrow \Omega_{\bar{V}/V} \otimes_{\bar{V}} \bar{R} \rightarrow \Omega_{\bar{R}/R} \rightarrow \Omega_{\bar{R}/\bar{R}\bar{V}} \rightarrow 0,$$

zusammen mit der Bestimmung des ersten Terms durch J. Fontaine, führt zu der gesuchten zweiten Abbildung, das heißt zu einer Beziehung zwischen Differentialen und Galois-Kohomologie.

In einigen speziellen Fällen kann man noch mehr aussagen: Man findet eine Beziehung zwischen der p -adischen étalen Kohomologie und der kristallinen Kohomologie, so daß die eine die andere eindeutig bestimmt. Dies ist ein Fall des "mysterious functor," nach dem schon A. Grothendieck gesucht hat. Der Beweis von J. Fontaine und W. Messing benutzt die "syntomic topology," und wird sicher in einem der Spezialvorträge ausführlicher dargestellt werden.

Außerdem sei erwähnt, daß Fontaine kürzlich eine weitere Vermutung von Schafarevitch bewiesen hat: Es gibt keine abelsche Varietät über \mathbb{Q} , welche

überall gute Reduktion hat. Der Beweis benutzt die Theorie der endlichen Gruppenschemata, und natürlich p -adische Darstellungen.

Literatur: [F5, F6, Fo].

4. Arithmetische Schnitt-Theorie. Beim Studium der Zahlkörper hat es sich gezeigt, daß man am besten versucht, die unendlichen Stellen und die endlichen ähnlich zu behandeln. Wenn man nun Varietäten über Zahlkörpern betrachtet, so ist es klar was man an den endlichen Stellen zu tun hat: Man benötigt gute Modelle über den entsprechenden diskreten Bewertungsringen. Was aber entspricht dem im Unendlichen? Die Erfahrung zeigt, daß man die algebraischen Objekte mit Metriken versehen muß. Wenn man zum Beispiel Geradenbündel auf einer algebraischen Kurve (über dem Zahlkörper) betrachtet, so benötigt man zunächst eine Ausdehnung auf ein Modell über den ganzen Zahlen. Dies erledigt die endlichen Stellen. Fürs Unendliche versteht man die Geradenbündel mit Metriken. Auf diese Weise hat S. Arakelov eine Schnitt-Theorie für arithmetische Flächen entwickelt, und man kann zeigen, daß ein Teil der Theorie der algebraischen Flächen sich auf diesen Fall überträgt. Dies gilt etwa für den Satz von Riemann-Roch oder für den Indexsatz von Hodge.

Man macht dabei die folgenden Überlegungen: Zunächst definiert man Volumenformen auf der Kohomologie eines metrisierten Geradenbündels. Dies ist ein Resultat über Riemann'sche Flächen, also rein lokal an den unendlichen Stellen. Die Definition des Volumens ist so gemacht, daß der übliche Beweis des Satzes von Riemann-Roch für algebraische Flächen übertragen werden kann. Aus dem Riemann-Roch leitet man eine Beziehung her zwischen Schnitt-Theorie und Néron-Tate Höhen, und der Satz von Hodge folgt. Außerdem gelten einige weitere Resultate, insbesondere ist bei Kurven vom Geschlecht größer als Eins das Quadrat des dualisierenden Geradenbündels ω kleiner oder gleich Null. Was aber etwa fehlt, ist ein Analogon zur Hodge-Theorie oder zum Frobenius.

Es stellt sich nun die Frage, was man bei höheren Dimensionen tun soll. Schon an den endlichen Stellen treten neue Schwierigkeiten auf, da die Theorie der minimalen Modelle fehlt, Trotzdem haben A. Beilinson, H. Gillet, und C. Soulé eine Schnitt-Theorie entwickelt. Es handelt sich um eine sehr schöne Mischung aus Analysis und Algebra, und meiner Ansicht nach liegt hier ein reiches Feld für künftige Forschungen.

Ein verwandtes Feld sind die Beilinson-Vermutungen über spezielle Werte von L -Reihen. Dabei scheint mir ein sehr wichtiger Punkt die Frage nach der Konstruktion von Motiven. Ihre Existenz wurde von A. Grothendieck vermutet, und diente als Leitmotiv für viele kohomologische Untersuchungen. Sie sollten eine Art von universeller Kohomologie-Theorie bilden. Zur Zeit kennen wir eigentlich nur 1-Motive, und dies sind im Wesentlichen semiabelsche algebraische Gruppen. Beilinson hat mit Hilfe der algebraischen K -Theorie einige Kandidaten für Motive höherer Dimension gefunden.

Literatur: [B1, B2, B3, F1, GS].

LITERATUR

- [B1] A. Beilinson, *Higher regulators and values of L-functions*, Modern Problems in Mathematics, VINITI Series **24** (1984), 181–238.
- [B2] —, *Notes on absolute Hodge cohomology*, preprint, 1984.
- [B3] —, *Height pairings between algebraic cycles*, preprint, 1985.
- [D] P. Deligne, *Seminaire Bourbaki* **616** (1984).
- [F1] G. Faltings, *Calculus on arithmetic surfaces*, Ann. of Math. **119** (1984), 387–424.
- [F2] —, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), 349–366.
- [F3] —, *Die Vermutungen von Tate und Mordell*, Jahresber. Deutsch. Math.-Verein. **86** (1984), 1–13.
- [F4] —, *Arithmetische Kompaktifizierung des Modulraums der abelschen Varietäten*, Lecture Notes in Math., vol. 1111, Springer-Verlag, 1985, pp. 321–383.
- [F5] —, *Hodge-Tate structures and modular forms*, submitted to Math. Ann.
- [F6] —, *P-adic Hodge-theory*, Manuscript, Princeton University, Princeton, N.J., 1985.
- [Fo] J. Fontaine, *Il n'y a pas de variété abélienne sur \mathbb{Z}* , Invent. Math. **81** (1985), 515–538.
- [FW] G. Faltings und G. Wüstholz, *Rational points*, Vieweg, Braunschweig, 1984.
- [G1] D. Goldfeld, *The class number of quadratic fields and the conjecture of Birch and Swinnerton-Dyer*, Ann. Scuola Norm. Sup. Pisa **3** (1976), 623–663.
- [G2] —, *Gauss' class number problem for imaginary quadratic fields*, Bull. Amer. Math. Soc. (N.S.) **13** (1985), 23–37.
- [GS] H. Gillet und C. Soulé, *Intersection sur les variétés d'Arakelov*, C. R. Acad. Sci. Paris **299** (1984), 563–566.
- [GZ] B. Gross und D. Zagier, *Points de Heegner et dérivés de fonctions L*, C. R. Acad. Sci. Paris Sér. I Math. **297** (1983), 85–87.
- [O] J. Osterlé, *Seminaire Bourbaki* **631** (1984).
- [S1] L. Szpiro, *Seminaire Bourbaki* **619** (1984).
- [S2] —, *Séminaire sur les pinceaux arithmétiques: La conjecture de Mordell*, Astérisque **127** (1985).
- [Z1] J. G. Zarhin, *Isogenies of abelian varieties over fields of finite characteristic*, Math. USSR-Sb. **24** (1974), 451–461.
- [Z2] —, *Endomorphisms of abelian varieties over fields of finite characteristic*, Math. USSR Izv. **9** (1975), 255–260.
- [Zg] D. Zagier, *Modular points, modular curves, modular surfaces and modular forms*, Lecture Notes in Math., vol. 1111, Springer-Verlag, pp. 225–248.

PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08544, USA

Topics in Quasiconformal Mappings

F. W. GEHRING

I. Introduction.

1. *Notation.* For $n \geq 1$ we let \mathbb{R}^n denote euclidean n -space, and for $x \in \mathbb{R}^n$ and $0 < r < \infty$ we let $\mathbb{B}^n(x, r)$ denote the open n -ball with center x and radius r , $\mathbb{S}^{n-1}(x, r) = \partial\mathbb{B}^n(x, r)$, $\mathbb{B}^n = \mathbb{B}^n(0, 1)$, and $\mathbb{S}^{n-1} = \mathbb{S}^{n-1}(0, 1)$. We also denote by $\bar{\mathbb{R}}^n = \mathbb{R}^n \cup \{\infty\}$ the one point compactification of \mathbb{R}^n equipped with the chordal metric

$$q(x, y) = |p(x) - p(y)|, \quad (1.1)$$

where p denotes stereographic projection of $\bar{\mathbb{R}}^n$ onto the sphere \mathbb{S}^n in \mathbb{R}^{n+1} . Throughout this paper, all notions of topology and convergence will be taken with respect to this metric.

Suppose that D and D' are domains in $\bar{\mathbb{R}}^n$ and that $f: D \rightarrow D'$ is a homeomorphism. We let

$$H_f(x) = \limsup_{r \rightarrow 0} H_f(x, r) \quad (1.2)$$

for $x \in D \setminus \{\infty, f^{-1}(\infty)\}$, where for $0 < r < \text{dist}(x, \partial D)$

$$H_f(x, r) = \frac{\max\{|f(x) - f(y)|: |x - y| = r\}}{\min\{|f(x) - f(z)|: |x - z| = r\}}, \quad (1.3)$$

and we extend $H_f(x)$ to the points ∞ and $f^{-1}(\infty)$ by setting $H_f(\infty) = H_{f \circ g}(0)$ and $H_f(f^{-1}(\infty)) = H_{g \circ f}(f^{-1}(\infty))$, where $g(x) = x/|x|^2$. When $n \geq 2$, we call

$$K(f) = \begin{cases} \infty & \text{if } \sup_{x \in D} H_f(x) = \infty, \\ \text{ess sup}_{x \in D} H_f(x) & \text{if } \sup_{x \in D} H_f(x) < \infty \end{cases} \quad (1.4)$$

the *linear dilatation* of f in D . For the purposes of this lecture, we say that f is *quasiconformal* if $K(f) < \infty$ and *K -quasiconformal* if $K(f) \leq K$, $1 \leq K < \infty$. Thus a homeomorphism is quasiconformal if it distorts the shape of an infinitesimal $(n-1)$ -sphere about each point by at most a bounded factor; it is K -quasiconformal if, in addition, this factor does not exceed K at almost every point.

The following result shows that the class of quasiconformal mappings is, as the name suggests, an extension of the family of conformal mappings.

This research was supported in part by grants from the National Science Foundation.

1.5. THEOREM. Suppose that D, D' are domains in $\bar{\mathbb{R}}^n$ and that $f: D \rightarrow D'$ is a homeomorphism. If $n = 2$, then f is 1-quasiconformal if and only if f or its complex conjugate is a meromorphic function of a complex variable in D . If $n \geq 3$, then f is 1-quasiconformal if and only if f is the restriction to D of a Möbius transformation, i.e., the composition of a finite number of reflections in $(n - 1)$ -spheres and planes.

When $n = 2$, Theorem 1.5 is simply a restatement of a theorem due to Mönchhoff [M3]. When $n \geq 3$, Theorem 1.5 is an extension of a well-known result of Liouville to a context which requires *no a priori* hypotheses on the smoothness of f [G3, R3].

2. *Historical remarks.* Plane quasiconformal mappings have been studied for almost sixty years. They appear in the late 1920s in papers by Grötzsch, who considered the problem of determining the *most nearly conformal* homeomorphisms between pairs of topologically equivalent plane configurations with one conformal invariant [G14]. They occur later under the name *quasiconformal* in a paper by Ahlfors on covering surfaces [A1].

In the late 1930s Teichmüller vastly extended the study of Grötzsch to mappings between closed Riemann surfaces and obtained a very natural parameter space for surfaces of fixed genus g , a space which is homeomorphic to \mathbb{R}^{6g-6} [T1]. At about the same time, Lavrentieff and Morrey generalized a classical result due to Gauss on the existence of isothermal coordinates by establishing versions of what is now known as the *measurable Riemann mapping theorem* for quasiconformal mappings [L1, M4].

In recent years, Ahlfors, Bers, and their school have greatly expanded the results of Teichmüller and applied plane quasiconformal mappings with success to a variety of areas in complex analysis, including kleinian groups and surface topology [A5, B6, E1, K2]. Sullivan's recent solution of the Fatou-Julia problem shows that this class can also be used very effectively to study problems on the iteration of rational functions [S3, S4].

Higher dimensional quasiconformal mappings were already considered by Lavrentieff in the 1930s [L2]. However, no systematic tool for studying this class was available until 1959 when Loewner introduced the notion of *conformal capacity* to show that \mathbb{R}^n cannot be mapped quasiconformally onto a proper subset of itself [L7].

Subsequently, Gehring, Väisälä, and many others applied Loewner's method and its equivalent extremal length formulation to develop the initial results for quasiconformal mappings in \mathbb{R}^n [G3, V1]. Then in the late 1960s, Reshetnyak and the Finnish school initiated a series of papers which extended the higher dimensional theory to noninjective quasiconformal, or *quasiregular*, mappings [M1, R2, V4], a study which recently resulted in Rickman's remarkable extension of the Picard theorem [R6].

3. *Role played by quasiconformal mappings.* Plane quasiconformal mappings constitute an important tool in complex analysis and they are particularly

valuable in the study of Riemann surfaces and discontinuous groups. Bers's theorem on simultaneous uniformization [B3] is a beautiful application of the measurable Riemann mapping theorem, while Drasin's solution of the inverse problem of Nevanlinna theory [D2] illustrates how this theorem can be used to attack problems of complex analysis in a manner similar to the way the $\bar{\partial}$ -equation has been applied in harmonic analysis and several complex variables.

The geometric proofs usually required to establish quasiconformal analogues of results for conformal mappings sometimes yield new insight into classical theorems and methods of complex function theory [L3]. Quasiconformal mappings also arise in exciting and unexpected ways in other parts of mathematics, for example, in harmonic analysis in connection with functions of bounded mean oscillation and singular integrals [B1], and in geometry and elasticity in connection with the injectivity and extension of quasi-isometries.

Higher dimensional quasiconformal mappings offer a new and nontrivial extension of complex analysis to \mathbb{R}^n which is distinct from [N2] and perhaps more geometric and flexible than the analytic theory through several complex variables. These mappings have been applied to solve problems in differential geometry, and they constitute a closed class of mappings, interpolating between homeomorphisms and diffeomorphisms, for which many results of geometric topology hold regardless of dimension. Finally, some of the methods developed to study higher dimensional quasiconformal mappings have found important applications in other branches of mathematics, for example, reverse Hölder inequalities in partial differential equations [G13].

4. *Comments on the above definition.* The quasiconformal mappings studied by Grötzsch and Teichmüller were assumed to be continuously differentiable at all but a finite number of points. Later Ahlfors [A2] and Bers [B2] observed that it was more natural to work with mappings $f: D \rightarrow D'$ for which one has the important inequality

$$K(f) \leq \liminf_{j \rightarrow \infty} K(f_j), \quad (4.1)$$

when $\{f_j\}$ is a sequence of homeomorphisms which converge to f locally uniformly in D . Indeed, we defined $K(f)$ as in (1.4), rather than by means of the simpler formula

$$K(f) = \sup_{x \in D} H_f(x), \quad (4.2)$$

just so that (4.1) would hold.

Inequality (4.1) implies that the class of K -quasiconformal mappings is closed with respect to locally uniform convergence. Moreover, when $n = 2$, the measurable Riemann mapping theorem implies that every homeomorphism f with $K(f) \leq K$ is the locally uniform limit of continuously differentiable homeomorphisms f_j with $K(f_j) \leq K$ [L3]. When $n = 3$, a quite different argument yields the same conclusion with $K(f_j) \leq \tilde{K}$ where \tilde{K} depends only on K [K1]. The situation when $n > 3$ appears to be open.

If $f: D \rightarrow D'$ is a homeomorphism with $K(f) < \infty$, then the Rademacher-Stepanoff theorem and an argument similar to that used by Menchoff imply that f is differentiable with Jacobian $J_f \neq 0$ a.e. in D , that f belongs to the Sobolev class $W_{1,\text{loc}}^n(D)$, and that $K(f^{-1}) = K(f)$ [G3]. Thus the inverse of a K -quasiconformal mapping is K -quasiconformal; similarly, the composition of a K_1 - and a K_2 -quasiconformal mapping is $K_1 K_2$ -quasiconformal. Though 1-quasiconformal mappings are real analytic, there exists for each $K > 1$ a K -quasiconformal self mapping f of \mathbb{R}^n which is not differentiable in a set of Hausdorff dimension n .

5. *Remark.* Since there are several excellent expository articles on plane quasiconformal mappings and their connections with Teichmüller spaces [A6, B4, B5, B7], the remainder of this lecture will emphasize the less developed theory in higher dimensions. In Chapter II we consider some basic results and open problems for quasiconformal mappings, comparing what is known for $n = 2$ and for $n > 2$. Then in Chapter III we mention several instances where these mappings arise naturally in other areas of mathematics.

II. Some results and open problems.

6. *Tools for studying quasiconformal mappings.* A homeomorphism $f: D \rightarrow D'$ is quasiconformal if the distortion function H_f in (1.2) is bounded. This is a local restriction and we must find some way to integrate it over D in order to obtain global properties of f . In classical complex function theory, this is accomplished by means of the Cauchy integral formula. Though Pompeiu's analogue is sometimes useful in treating plane quasiconformal mappings, the tool most often used to replace the Cauchy formula is the method of extremal length, formulated by Ahlfors and Beurling [A8], and its extension to higher dimensions [F3, V3].

7. *Modulus of a curve family.* Suppose that Γ is a family of curves in $\bar{\mathbb{R}}^n$ and let $\text{adm}(\Gamma)$ denote the collection of all Borel measurable functions $\rho: \mathbb{R}^n \rightarrow [0, \infty]$ such that $\int_\gamma \rho \, ds \geq 1$ for each locally rectifiable curve γ in Γ . Then

$$\text{mod}(\Gamma) = \inf_{\rho \in \text{adm}(\Gamma)} \int_{\mathbb{R}^n} \rho^n \, dm \quad \text{and} \quad \lambda(\Gamma) = \text{mod}(\Gamma)^{1/(1-n)} \quad (7.1)$$

are the *conformal modulus* and *extremal length*, respectively, of Γ .

It is not difficult to see that $\text{mod}(\Gamma)$ is an outer measure on the space of all curve families in $\bar{\mathbb{R}}^n$. Alternatively, if we regard the curves in Γ as homogeneous wires, then we may think of $\lambda(\Gamma)$ as the resistance of the family Γ . In particular, $\text{mod}(\Gamma)$ is large if the curves in Γ are short and plentiful, and small otherwise.

The importance of the conformal modulus in the present context is due to its quasi-invariance with respect to quasiconformal mappings.

7.2. THEOREM. *If $f: D \rightarrow D'$ is K -quasiconformal and if Γ is a family of curves which lie in D , then*

$$K^{1-n} \text{mod}(\Gamma) \leq \text{mod}(f(\Gamma)) \leq K^{n-1} \text{mod}(\Gamma). \quad (7.3)$$

Inequality (7.3) plays a key role in the study of quasiconformal mappings. For this reason it is customary to refer to

$$K^*(f) = \max \left(\sup_{\Gamma} \left(\frac{\text{mod}(f(\Gamma))}{\text{mod}(\Gamma)} \right), \sup_{\Gamma} \left(\frac{\text{mod}(\Gamma)}{\text{mod}(f(\Gamma))} \right) \right) \quad (7.4)$$

as the *maximal dilatation* of f and say that f is K -*quasiconformal* if $K^*(f) \leq K$; here the supremum in (7.4) is taken over all curve families Γ in D for which $\text{mod}(\Gamma)$ and $\text{mod}(f(\Gamma))$ are not simultaneously 0 or ∞ . The inequality

$$K(f)^{n/2} \leq K^*(f) \leq K(f)^{n-1} \quad (7.5)$$

shows that this definition yields the same class of quasiconformal mappings and that $K^*(f) = K(f)$ whenever $n = 2$ or $K(f) = 1$.

A homeomorphism $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is quasiconformal if and only if there exists a constant c such that

$$\limsup_{r \rightarrow 0} H_f(x, r) \leq c \quad (7.6)$$

for all $x \in \mathbb{R}^n$. We illustrate the use of (7.3) by establishing a global form of this inequality.

7.7. THEOREM. *If $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is K -quasiconformal, then*

$$H_f(x, r) \leq c \quad (7.8)$$

for all $x \in \mathbb{R}^n$ and $0 < r < \infty$, where $c = c(K, n)$.

The proof depends on two estimates for the conformal moduli of certain curve families [G2, G12, V1].

7.9. LEMMA. *If $0 < a < b < \infty$ and if Γ is a family of open arcs in $\bar{\mathbb{R}}^n$ which join $\mathbb{S}^{n-1}(0, a)$ to $\mathbb{S}^{n-1}(0, b)$, then*

$$\text{mod}(\Gamma) \leq \omega_{n-1}(\log \frac{b}{a})^{1-n},$$

where ω_{n-1} denotes the $(n-1)$ -measure of \mathbb{S}^{n-1} .

7.10. LEMMA. *If C_1 and C_2 are disjoint continua in $\bar{\mathbb{R}}^n$ which join 0 to $\mathbb{S}^{n-1}(0, a)$ and ∞ to $\mathbb{S}^{n-1}(0, b)$, respectively, and if Γ is the family of all open arcs which join C_1 to C_2 in $\bar{\mathbb{R}}^n \setminus (C_1 \cup C_2)$, then*

$$\text{mod}(\Gamma) \geq \omega_{n-1}(\log(\lambda_n(\frac{b}{a} + 1)))^{1-n},$$

where λ_n depends only on n .

7.11. COROLLARY. *If $n > 2$, if C_1 and C_2 are disjoint, linked continua in $\bar{\mathbb{R}}^n$ and if Γ is the family of all open arcs which join C_1 and C_2 in $\bar{\mathbb{R}}^n \setminus (C_1 \cup C_2)$, then $\text{mod}(\Gamma) \geq c$ where $c = c(n) > 0$.*

PROOF OF THEOREM 7.7. By performing preliminary translations, we may assume that $x = 0$ and $f(0) = 0$. Let m and M denote the minimum and maximum values assumed by $|f|$ on $\mathbb{S}^{n-1}(0, r)$ and suppose that $m < M$. Next set

$$C_1 = \{x \in \mathbb{R}^n: |f(x)| \leq m\}, \quad C_2 = \{x \in \mathbb{R}^n: |f(x)| \geq M\} \cup \{\infty\},$$

and let Γ be the family of open arcs which join C_1 and C_2 in $\bar{\mathbb{R}}^n \setminus (C_1 \cup C_2)$. Then the above estimates and (7.3) imply that

$$\omega_{n-1}(\log 2\lambda_n)^{1-n} \leq \text{mod}(\Gamma) \leq K^{n-1} \text{mod}(f(\Gamma)) \leq K^{n-1} \omega_{n-1}(\log(M/m))^{1-n}$$

and we obtain (7.8) with $c = (2\lambda_n)^K$.

8. Mapping problems. A basic question in this area is to decide when two domains in $\bar{\mathbb{R}}^n$ are quasiconformally equivalent, i.e., if one can be mapped quasiconformally onto the other. Since the general case is quite difficult even when $n = 2$, we consider here the simpler problem of characterizing the domains D in $\bar{\mathbb{R}}^n$ which are quasiconformally equivalent to the unit ball \mathbb{B}^n . The Riemann mapping theorem and the estimates in Lemmas 7.9 and 7.10 yield a complete answer when $n = 2$.

8.1. THEOREM. *A domain D in $\bar{\mathbb{R}}^2$ is quasiconformally equivalent to \mathbb{B}^2 if and only if ∂D is a nondegenerate continuum.*

No such characterization exists in higher dimensions. Indeed, the domains D_3 and D_4 in (8.6) below show that when $n > 2$, there is no way to decide whether the image of \mathbb{B}^n under a self homeomorphism of $\bar{\mathbb{R}}^n$ is quasiconformally equivalent to \mathbb{B}^n by looking only at its boundary.

The following sufficient condition is a consequence of methods used to treat the higher dimensional Schoenflies problem [G5, M2].

8.2. THEOREM. *A domain D in $\bar{\mathbb{R}}^n$ is quasiconformally equivalent to \mathbb{B}^n if there exist closed sets $E \subset D$, $E' \subset \mathbb{B}^n$ and a quasiconformal mapping $g: D \setminus E \rightarrow \mathbb{B}^n \setminus E'$ such that $|g(x)| \rightarrow 1$ as $x \rightarrow \partial D$ in D .*

As in the topological case, localized versions of Theorem 8.2 can be established when D is a Jordan domain in $\bar{\mathbb{R}}^n$, i.e., when ∂D is homeomorphic to \mathbb{S}^{n-1} [B10, G1].

8.3. COROLLARY. *If D is a domain in \mathbb{R}^n and if D is diffeomorphic to \mathbb{S}^{n-1} , then D is quasiconformally equivalent to \mathbb{B}^n .*

It is easy to construct a domain in \mathbb{R}^n which is quasiconformally equivalent to \mathbb{B}^n and does not have a tangent plane at any point of its boundary [G12]. Thus the sufficient condition in Corollary 8.3 is far from necessary.

A necessary condition for quasiconformal equivalence to \mathbb{B}^n depends on the following refinement of the notion of local connectivity. A set $E \subset \bar{\mathbb{R}}^n$ is said to be *linearly locally connected* if there exists a constant c , $1 \leq c < \infty$, such that for each $x \in \mathbb{R}^n$ and $0 < r < \infty$

$$\begin{aligned} E \cap \bar{\mathbb{B}}^n(x, r) & \text{ lies in a component of } E \cap \bar{\mathbb{B}}^n(x, cr), \\ E \setminus \mathbb{B}^n(x, r) & \text{ lies in a component of } E \setminus \mathbb{B}^n(x, r/c). \end{aligned} \tag{8.4}$$

Then an argument based again on inequality (7.3) and the estimates in Lemma 7.9 and Corollary 7.11 implies the following result [G7, G12].

8.5. THEOREM. *If $n > 2$ and if D in $\bar{\mathbb{R}}^n$ is quasiconformally equivalent to \mathbb{B}^n , then $\bar{\mathbb{R}}^n \setminus D$ is linearly locally connected.*

Theorem 8.5 yields many simple domains in \mathbb{R}^n which are homeomorphic, but not quasiconformally equivalent, to \mathbb{B}^n . For example, let

$$\begin{aligned} D_1 &= \{x \in \mathbb{R}^n : r < 1, |x_n| < \infty\}, \\ D_2 &= \{x \in \mathbb{R}^n : r < \infty, |x_n| < 1\}, \\ D_3 &= \{x \in \mathbb{R}^n : x_n > \min(r^{1/2}, 1)\}, \\ D_4 &= \{x \in \mathbb{R}^n : x_n < \min(r^{1/2}, 1)\}, \end{aligned} \tag{8.6}$$

where $x = (x_1, \dots, x_n)$ and $r = (x_1^2 + \dots + x_{n-1}^2)^{1/2}$. Then explicit constructions yield homeomorphisms which map D_1 and D_3 quasiconformally onto \mathbb{B}^n . On the other hand when $n > 2$, $\bar{\mathbb{R}}^n \setminus D_2$ and $\bar{\mathbb{R}}^n \setminus D_4$ are not linearly locally connected and hence D_2 and D_4 are not quasiconformally equivalent to \mathbb{B}^n .

The necessary condition in Theorem 8.5 is not sufficient and the problem of finding sharp geometric criteria for testing quasiconformal equivalence to \mathbb{B}^n remains a most interesting open question.

9. *Homeomorphic and quasiconformal extensions.* Suppose that D and D' are domains in $\bar{\mathbb{R}}^n$ and that $f: D \rightarrow D'$ is quasiconformal. We consider next under what circumstances f admits a homeomorphic extension to \bar{D} or a quasiconformal extension to $\bar{\mathbb{R}}^n$.

9.1. THEOREM. *If D and D' are simply-connected domains of hyperbolic type in $\bar{\mathbb{R}}^2$, then each quasiconformal $f: D \rightarrow D'$ has a homeomorphic extension to \bar{D} if and only if D and D' are Jordan domains.*

The sufficiency in Theorem 9.1 follows from a theorem of Ahlfors [A2] and the necessity from [E3]. In higher dimensions we have the following result [V2].

9.2. THEOREM. *If D and D' are Jordan domains in $\bar{\mathbb{R}}^n$ and if D is quasiconformally equivalent to \mathbb{B}^n , then each quasiconformal $f: D \rightarrow D'$ has a homeomorphic extension to \bar{D} .*

When $n = 2$, the second hypothesis in Theorem 9.2 is superfluous since every Jordan domain is conformally equivalent to \mathbb{B}^2 . When $n > 2$, this is not the case as seen by the examples in (8.6), and Theorem 9.2 does not hold without this additional restriction [K3].

As to the problem of quasiconformal extension to $\bar{\mathbb{R}}^n$, we say that a set E in $\bar{\mathbb{R}}^2$ is a K -quasidisk or K -quasicircle if it is the image of \mathbb{B}^2 or \mathbb{S}^1 , respectively, under a K -quasiconformal self mapping of $\bar{\mathbb{R}}^2$. By a theorem of Ahlfors, a Jordan domain D is a quasidisk if and only if there exists a constant c such that

$$\min_{j=1,2} \text{dia}(\gamma_j) \leq c|z_1 - z_2| \tag{9.3}$$

for each $z_1, z_2 \in \partial D$, where γ_1 and γ_2 denote the components of $\partial D \setminus \{z_1, z_2\}$ [A3].

9.4. THEOREM. *If D and D' are Jordan domains in $\bar{\mathbb{R}}^2$, then each quasiconformal $f: D \rightarrow D'$ has a quasiconformal extension to $\bar{\mathbb{R}}^2$ if and only if D and D' are quasidisks.*

A simply-connected domain D in $\bar{\mathbb{R}}^2$ is a quasidisk if and only if it is linearly locally connected. Hence this notion also arises in connection with quasiconformal extension.

The sufficiency in Theorem 9.4 is due to Ahlfors [A2] and the necessity to Rickman [R5]. A higher dimensional analogue of this result is as follows [G4, V5].

9.5. THEOREM. *If $n > 2$ and if D is a Jordan domain in $\bar{\mathbb{R}}^n$, then each quasiconformal $f: D \rightarrow \mathbb{B}^n$ has a quasiconformal extension to $\bar{\mathbb{R}}^n$ if and only if $D^* = \bar{\mathbb{R}}^n \setminus \bar{D}$ is quasiconformally equivalent to \mathbb{B}^n .*

Thus the problem of quasiconformal extension in higher dimensions differs from the plane case in two respects. First, when $n = 2$, the exterior D^* of every Jordan domain D is quasiconformally equivalent to \mathbb{B}^2 ; this is not true when $n > 2$ as we observed above. Second, when $n > 2$, each quasiconformal $f: D \rightarrow \mathbb{B}^n$ has a quasiconformal extension to $\bar{\mathbb{R}}^n$ whenever D^* is quasiconformally equivalent to \mathbb{B}^n ; this is not true when $n = 2$ since there exist Jordan domains D which do not satisfy condition (9.3) and hence are not quasidisks.

10. *Boundary correspondence.* We turn to the problem of characterizing the boundary mappings induced by quasiconformal self mappings of balls and half-spaces. For $n \geq 2$ let \mathbb{H}^n denote the upper halfspace $\{x \in \mathbb{R}^n: x_n > 0\}$. Then each quasiconformal $f: \mathbb{H}^n \rightarrow \mathbb{H}^n$ has a quasiconformal extension \tilde{f} to $\bar{\mathbb{R}}^n$ whose restriction to $\partial\mathbb{H}^n$ is a self homeomorphism φ of $\bar{\mathbb{R}}^{n-1}$. The problem of studying such boundary correspondences was initiated by Beurling and Ahlfors [B8].

10.1. THEOREM. *A homeomorphism $\varphi: \bar{\mathbb{R}}^1 \rightarrow \bar{\mathbb{R}}^1$ with $\varphi(\infty) = \infty$ is the boundary correspondence for a quasiconformal self mapping f of \mathbb{H}^2 with $\tilde{f}(\infty) = \infty$ if and only if there exists a constant c such that*

$$\frac{1}{c} \leq \frac{\varphi(x+r) - \varphi(x)}{\varphi(x) - \varphi(x-r)} \leq c \quad (10.2)$$

for all $x \in \mathbb{R}^1$ and $0 < r < \infty$.

Inequality (10.2) is equivalent to the requirement that $H_\varphi(x, r) \leq c$. This condition is replaced by its local form $H_\varphi(x) \leq c$, or that φ is quasiconformal, in the higher dimensional analogue of Theorem 10.1.

10.3. THEOREM. *When $n > 2$, a homeomorphism $\varphi: \bar{\mathbb{R}}^{n-1} \rightarrow \bar{\mathbb{R}}^{n-1}$ is the boundary correspondence of a quasiconformal self mapping f of \mathbb{H}^n if and only if φ is itself quasiconformal.*

The necessity in Theorems 10.1 and 10.3 follows, respectively, from inequalities (7.8) and (7.6) and the fact that

$$H_\varphi(x, r) \leq H_{\tilde{f}}(x, r) \quad \text{and} \quad H_\varphi(x) \leq H_{\tilde{f}}(x) \quad (10.4)$$

for relevant x and r .

Beurling and Ahlfors established the sufficiency in Theorem 10.1 by showing that the formula

$$\begin{aligned} f(x_1, x_2) = & \frac{1}{2x_2} \int_0^{x_2} (\varphi(x_1 + t) + \varphi(x_1 - t)) dt \\ & + \frac{i}{2x_2} \int_0^{x_2} (\varphi(x_1 + t) - \varphi(x_1 - t)) dt \end{aligned} \quad (10.5)$$

defines a quasiconformal extension of φ to \mathbb{H}^2 .

Ahlfors [A4] modified this construction and used the fact that each quasiconformal $\varphi: \bar{\mathbb{R}}^2 \rightarrow \bar{\mathbb{R}}^2$ can be written as the composition of mappings with small dilatation (see Corollary 11.4) to obtain a quasiconformal extension of φ to \mathbb{H}^3 and thus prove the sufficiency in Theorem 10.3 when $n = 3$. Next Carleson [C1] employed quite different methods from three-dimensional topology to extend each quasiconformal $\varphi: \bar{\mathbb{R}}^3 \rightarrow \bar{\mathbb{R}}^3$ to \mathbb{H}^4 . Finally, Tukia and Väisälä [T5] started from an idea of Carleson's and applied results of Sullivan's [S1] to establish the sufficiency in Theorem 10.3 for general n .

After composition with suitable Möbius transformations, (10.2) yields a cross ratio characterization for the boundary mappings $\varphi: \partial D \rightarrow \partial D$ induced by arbitrary quasiconformal self mappings of a disk or halfplane D in \mathbb{R}^2 , and (10.5) gives an explicit quasiconformal extension $T(\varphi): D \rightarrow D$ of each such correspondence φ . Tukia [T4] recently settled an important problem in Teichmüller theory by showing that if G is a subgroup of $\text{Möb}(D)$, the group of all Möbius self mappings of D , then each G -compatible boundary correspondence $\varphi: \partial D \rightarrow \partial D$ has a G -compatible quasiconformal extension to D . Douady and Earle [D1] extended this work by exhibiting a *conformally natural* quasiconformal extension operator T_0 such that

$$g \circ T_0(\varphi) \circ h = T_0(g \circ \varphi \circ h) \quad (10.6)$$

for each homeomorphism $\varphi: \partial D \rightarrow \partial D$ and all $g, h \in \text{Möb}(D)$. This beautiful operator should yield many new results in the area; see [E2].

If D is a ball or halfspace in \mathbb{R}^n where $n > 2$, then the method of Douady and Earle assigns to each homeomorphism $\varphi: \partial D \rightarrow \partial D$ a continuous extension $T_0(\varphi): D \rightarrow D$ for which (10.6) holds. However, $T_0(\varphi)$ will, in general, be neither quasiconformal nor injective except when $K(\varphi)$ is small, i.e., $K(\varphi) \leq K_n$ where K_n depends only on n . It would be interesting to know if every quasiconformal φ has a conformally natural quasiconformal extension.

11. Measurable Riemann mapping theorem and decomposition. If $f: D \rightarrow D'$ is quasiconformal, then f has a nonsingular differential $df: \mathbb{R}^n \rightarrow \mathbb{R}^n$ at almost all $x \in D$. At each such x , $df = df(x)$ maps an ellipsoid $E_f = E_f(x)$ about 0 with minimum axis length 1 onto an $(n-1)$ -sphere about 0. Then $H_f(x)$ is the maximum axis length of $E_f(x)$ and the maximum stretching under f at x occurs in the directions of the smallest axes of $E_f(x)$. If $g: D' \rightarrow D''$ is quasiconformal, then g is conformal if and only if $E_{g \circ f} = E_f$ a.e. in D by Theorem 1.5, and E_f determines f up to postcomposition with a conformal mapping.

When $n = 2$ and f is sense preserving, E_f is determined by the *Beltrami coefficient* or *complex dilatation*

$$\mu_f(x) = f_{\bar{x}}/f_x, \quad x = x_1 + ix_2, \quad (11.1)$$

of f at x . In particular, μ_f is measurable with

$$|\mu_f(x)| = \frac{H_f(x) - 1}{H_f(x) + 1}, \quad \|\mu_f\|_{L^\infty} = \frac{K(f) - 1}{K(f) + 1} < 1, \quad (11.2)$$

and $\mu_{g \circ f} = \mu_f$ a.e. in D if and only if $g: D' \rightarrow D''$ is conformal. Moreover, in dimension two it is possible to prescribe the dilatation μ_f , and hence the ellipse E_f , at almost every $x \in D$ [A7].

11.3. MEASURABLE RIEMANN MAPPING THEOREM. *If μ is measurable with $\|\mu\|_{L^\infty} < 1$ in $\bar{\mathbb{R}}^2$, then there exists a quasiconformal self mapping $f = f_\mu$ of $\bar{\mathbb{R}}^2$ with $\mu_f = \mu$ a.e. If f is normalized to fix three points, then f is unique and depends holomorphically on μ .*

Theorem 11.3 is of fundamental importance in studying the complex structure on Teichmüller space. It can also be a powerful tool for attacking other problems of complex analysis. One example is the solution of the inverse problem of Nevanlinna theory [D2] where Drasin first constructed a locally quasiconformal function g with prescribed defects, and then applied Theorem 11.3 to obtain a quasiconformal self mapping f of $\bar{\mathbb{R}}^2$ so that $g \circ f$ was meromorphic with the same defects as g . A second example is Sullivan's recent solution [S3] of the Fatou-Julia problem on wandering domains where Theorem 11.3 was used to construct a large real analytic family of quasiconformal deformations of a given rational function.

The following is an important consequence of Theorem 11.3.

11.4. COROLLARY. *If $n = 2$ and $\varepsilon > 0$, then each quasiconformal $f: D \rightarrow D'$ can be written in the form $f = f_1 \circ \cdots \circ f_m$ where $K(f_j) < 1 + \varepsilon$ for $j = 1, \dots, m$ and $m = m(\varepsilon, K(f))$.*

There is no analogue of Theorem 11.3 in higher dimensions. Moreover, when $n > 2$, examples show that Corollary 11.4 is almost certainly not true without further restrictions on the domain D . It is an important open problem to decide if some higher dimensional form of this result holds, even for the case where $D = D' = \bar{\mathbb{R}}^n$.

12. Quasiconformal groups. A group G of self homeomorphisms of $\bar{\mathbb{R}}^n$ is said to be *discrete* if G contains no sequence of elements which converge to the identity uniformly in $\bar{\mathbb{R}}^n$, and *K -quasiconformal* if $K(g) \leq K$ for each g in G . Though the family of quasiconformal groups contains all Möbius groups, Theorem 11.3 can be used to show that this larger family does not exhibit new phenomena when $n = 2$ [S2, T2].

12.1. THEOREM. *When $n = 2$, each quasiconformal group G can be written in the form $G = f^{-1} \circ H \circ f$, where H is a Möbius group and f a quasiconformal self mapping of $\bar{\mathbb{R}}^2$.*

The situation is different in higher dimensions, and for each $n > 2$ there exists a quasiconformal group which is not even isomorphic as a topological group to a Möbius group [T3]. Nevertheless, the following convergence property allows one to establish quasiconformal analogues of many basic properties of Möbius groups [G10].

12.2. THEOREM. *If G is a discrete quasiconformal group, then for each infinite sequence of distinct elements in G there exists a subsequence $\{g_j\}$ and points x_0, y_0 in $\bar{\mathbb{R}}^n$ such that $g_j \rightarrow y_0$ locally uniformly in $\bar{\mathbb{R}}^n \setminus \{x_0\}$ and $g_j^{-1} \rightarrow x_0$ locally uniformly in $\bar{\mathbb{R}}^n \setminus \{y_0\}$.*

Suppose that G is a group of self homeomorphisms of $\bar{\mathbb{R}}^n$. We say that G is a *discrete convergence group* if it satisfies the conclusion of Theorem 12.2, and that an element g of G is *elliptic* if it is of finite order or periodic, and *parabolic* or *loxodromic* if it has infinite order and one or two fixed points, respectively. The *limit set* $L(G)$ is the complement of the ordinary set $O(G)$, the set of $x \in \bar{\mathbb{R}}^n$ which have a neighborhood U such that $g(U) \cap U \neq \emptyset$ for at most finitely many $g \in G$. Finally, G is *properly discontinuous* in an open set O if for each compact $F \subset O$, $g(F) \cap F \neq \emptyset$ for at most finitely many $g \in G$ [G10].

12.3. THEOREM. *Suppose that G is a discrete convergence group. Then each element of G is elliptic, parabolic, or loxodromic, and the limit set $L(G)$ is nowhere dense or equal to $\bar{\mathbb{R}}^n$. Moreover if $\text{card}(L(G)) > 2$, then $L(G)$ is perfect, $L(G)$ lies in the closure of each nonempty G -invariant set, and the set of fixed point pairs of loxodromic elements in G is dense in $L(G) \times L(G)$.*

Though discrete convergence groups resemble Möbius groups in many respects, examples exist which show that they need not be topologically conjugate to Möbius groups [F2, G10]. They also occur quite naturally in situations which have nothing to do with Möbius or quasiconformal groups.

12.4. THEOREM. *A group G of self homeomorphisms of $\bar{\mathbb{R}}^n$ is a discrete convergence group if it is properly discontinuous in $\bar{\mathbb{R}}^n \setminus E$, where E is closed and totally disconnected.*

It will be interesting to see how much of the classical theory of kleinian groups carries over for this general class of groups.

13. Hölder continuity and integrability. Theorem 12.2 can be deduced from (4.1) and the following estimate for change in the chordal distance q in (1.1) under a quasiconformal mapping [G7].

13.1. THEOREM. *If $f: D \rightarrow D'$ is K -quasiconformal and if $\bar{\mathbb{R}}^n \setminus D \neq \emptyset$, then*

$$q(f(x), f(y))q(\bar{\mathbb{R}}^n \setminus D') \leq c(q(x, y)/q(x, \partial D))^{1/K} \quad (13.2)$$

for x, y in D , where $q(E)$ denotes the chordal diameter of E and $c = c(n)$.

Theorem 13.1 is a consequence of (7.3) and Lemmas 7.9 and 7.10. When $D, D' \subset \mathbb{R}^n$, it implies that each K -quasiconformal $f: D \rightarrow D'$ is locally Hölder

continuous with exponent $1/K$ and hence that these mappings interpolate between diffeomorphisms and homeomorphisms for $1 \leq K < \infty$. This fact is also reflected in the integrability of the Jacobian J_f of f .

13.3. THEOREM. *If $f: D \rightarrow D'$ is K -quasiconformal where $D, D' \subset \mathbb{R}^n$, then J_f is locally L^p -integrable in D for $1 \leq p < p(K, n)$, where*

$$p(K, n) \leq K/(K-1), \quad \lim_{K \rightarrow 1} p(K, n) = \infty. \quad (13.4)$$

Bojarski [B9] established the existence of the exponent $p(K, n)$ for $n = 2$ by applying the Calderón-Zygmund inequality to the Beurling transform in (14.7). The proof for $n > 2$ was based on the fact that $g = |J_f|$ satisfies the reverse Hölder inequality

$$\frac{1}{m(Q)} \int_Q g \, dm \leq c \left(\frac{1}{m(Q)} \int_Q g^{1/n} \, dm \right)^n, \quad c = c(K, n), \quad (13.5)$$

for each n -cube Q in D with $\text{dia}(f(Q)) < d(f(Q), \partial D')$, and on a lemma to the effect that if (13.5) holds for all n -cubes Q contained in an n -cube Q' , then g belongs to $L^p(Q')$ for $1 \leq p < p(c, n)$ [G6]. These results were sharpened in [I2, R4] to obtain the second part of (13.4); the example

$$f(x) = |x|^{-a}x, \quad a = (K-1)/K, \quad (13.6)$$

gives the first part of (13.4). There is reason to suspect that one can take $p(K, n) = K/(K-1)$ in Theorem 13.3. However, this has not been established even for the case $n = 2$.

III. Connections with other areas of mathematics.

14. Harmonic and functional analysis. Quasiconformal mappings are encountered in harmonic analysis through their connections with functions of bounded mean oscillation and singular integrals.

A function u is said to be of *bounded mean oscillation* in a domain $D \subset \mathbb{R}^n$, or in $\text{BMO}(D)$, if u is locally integrable and

$$\|u\|_{\text{BMO}(D)} = \sup_B \frac{1}{m(B)} \int_B |u - u_B| \, dm < \infty, \quad (14.1)$$

where the supremum is taken over all n -balls B with $\bar{B} \subset D$ and

$$u_B = \frac{1}{m(B)} \int_B u \, dm. \quad (14.2)$$

The class BMO was introduced by John and Nirenberg [J3] in connection with John's work in elasticity [J1], and it gained great prominence when Fefferman showed that $\text{BMO}(\mathbb{R}^n)$ is the dual of the Hardy space $H^1(\mathbb{R}^n)$ [F1].

The following relations between the class BMO and quasiconformal mappings are due to Reimann [R1].

14.3. THEOREM. *If $f: D \rightarrow D'$ is K -quasiconformal where $D, D' \subset \mathbb{R}^n$, then $\|\log J_f\|_{\text{BMO}(D)} \leq c$ where $c = c(K, n)$.*

14.4. THEOREM. *Suppose that $f: D \rightarrow D'$ is a homeomorphism where $D, D' \subset \mathbb{R}^n$. Then f is quasiconformal if and only if there exists a constant c such that*

$$\frac{1}{c} \|u\|_{\text{BMO}(G')} \leq \|u \circ f\|_{\text{BMO}(G)} \leq c \|u\|_{\text{BMO}(G')} \quad (14.5)$$

for each subdomain G of D and each u continuous in $G' = f(G)$.

Theorem 14.3 and the necessity in Theorem 14.4 follow from the fact that $g = |J_f|$ satisfies the reverse Hölder inequality in (13.5). The sufficiency in Theorem 14.4 is a variant, due to Astala [A9], of Reimann's original result.

Theorem 14.4 characterizes quasiconformal mappings as the homeomorphisms which preserve the class BMO. The following result [J4] characterizes quasidisks in terms of extension properties for BMO.

14.6. THEOREM. *If D is a simply-connected domain of hyperbolic type in \mathbb{R}^2 , then each function u in $\text{BMO}(D)$ has a BMO extension to \mathbb{R}^2 if and only if D is a quasidisk.*

Next the best possible exponents for Jacobian integrability and area distortion for plane quasiconformal mappings are closely connected with sharp constants in two inequalities for the Beurling transform

$$Tg(x) = -\frac{1}{\pi} \int_{\mathbb{R}^2} \frac{g(y)}{(x-y)^2} dm. \quad (14.7)$$

For example, T is a bounded operator on $L^p(\mathbb{R}^2)$ with

$$\|T\|_p = \sup_g \frac{\|Tg\|_{L^p(\mathbb{R}^2)}}{\|g\|_{L^p(\mathbb{R}^2)}} \geq \max\left(p-1, \frac{1}{p-1}\right) \quad (14.8)$$

for $1 < p < \infty$ and $\|T\|_2 = 1$, and there is reason to believe that

$$\liminf_{p \rightarrow \infty} \frac{1}{p} \|T\|_p = 1. \quad (14.9)$$

If true, this would yield the sharp upper bound $p(K, 2) = K/(K-1)$ for the integrability of the Jacobian of a plane quasiconformal mapping discussed in §13 [I1].

Next, one can show that there exist constants a and b such that

$$\int_{\mathbb{B}^2} |T\chi_E(x)| dm \leq am(E) \log(\pi/m(E)) + bm(E) \quad (14.10)$$

for each measurable $E \subset \mathbb{B}^2$. This inequality can be combined with Theorem 11.3 to prove that

$$\frac{m(f(E))}{\pi} \leq c \left(\frac{m(E)}{\pi} \right)^{K^{-a}}, \quad (14.11)$$

$c = c(K) = 1 + O(K-1)$ as $K \rightarrow 1$, for each K -quasiconformal $f: \mathbb{B}^2 \rightarrow \mathbb{B}^2$ with $f(0) = 0$ and each measurable set $E \subset \mathbb{B}^2$ [G11]. Moreover, the above

reasoning can be reversed to show that if (14.11) holds for a given constant a , then so does (14.10). It is conjectured that both hold with $a = 1$. If so, this would again imply that $p(K, 2) = K/(K - 1)$.

Finally, the problem of quasiconformal equivalence of domains can be reformulated in terms of function algebras. Given a domain D in \mathbb{R}^n , we let $A(D)$ denote the algebra of functions $u \in C(D) \cap W_n^1(D)$ with norm

$$\|u\| = \|u\|_{L^\infty(D)} + \|\nabla u\|_{L^n(D)}, \quad (14.12)$$

the so-called *Royden algebra* of D . We then have the following result [L5, L6].

14.13. THEOREM. *Two domains D and D' in \mathbb{R}^n are quasiconformally equivalent if and only if $A(D)$ and $A(D')$ are isomorphic as algebras.*

Little is known about the structure of these algebras and it may be that geometric methods used to determine quasiconformal equivalence will yield more information about them than vice versa.

15. Quasi-isometries and elasticity. A mapping $f: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an *L -quasi-isometry* in E if

$$\frac{1}{L}|x_1 - x_2| \leq |f(x_1) - f(x_2)| \leq L|x_1 - x_2| \quad (15.1)$$

for $x_1, x_2 \in E$; f is a *local L -quasi-isometry* in E if for each $L' > L$, each $x \in E$ has a neighborhood U such that f is an L' -quasi-isometry in $E \cap U$.

If f is quasi-isometric in a domain D , then f is quasiconformal by (1.2) and (1.4); the mapping in (13.6) shows that the converse is false. Nevertheless, quasiconformal homeomorphisms arise in questions concerning extension and injectivity of these mappings.

15.2. THEOREM. *If $n \neq 4$, then a quasi-isometry f of E has a quasi-isometric extension to \mathbb{R}^n if and only if f has a quasiconformal extension to \mathbb{R}^n .*

Theorem 15.2 [T6] gives a criterion for extension in terms of the mapping f . There is also a criterion in terms of the set E when E is a Jordan curve [G9].

15.3. THEOREM. *If C is a Jordan curve in \mathbb{R}^2 , then each quasi-isometry f of C has a quasi-isometric extension to \mathbb{R}^2 if and only if C is a quasicircle.*

For each domain $D \subset \mathbb{R}^n$ let $L(D)$ denote the supremum of the numbers $L \geq 1$ with the property that each local L -quasi-isometry f in D is injective there. The constant $L(D)$ has a physical interpretation if we think of D as an elastic body and f as the deformation experienced by D when subjected to a force field. Requiring that f be a local L -quasi-isometry bounds the strain in D under the force field and $L(D)$ measures the critical strain in D before D collapses onto itself.

Little is known about this constant except that $2^{1/4} \leq L(D) \leq 2^{1/2}$ whenever D is a ball or halfspace [J2]. However, we can characterize a large class of plane domains for which $L(D) > 1$ [G8].

15.4. THEOREM. *If D is a simply-connected proper subdomain of \mathbb{R}^2 , then $L(D) > 1$ if and only if D is a quasidisk.*

15.5. COROLLARY. *If f is a local L -quasi-isometry of a bounded simply-connected domain in \mathbb{R}^2 and if $L < L(D)$, then f has an M -quasi-isometric extension to \mathbb{R}^2 where $M = M(L, L(D))$.*

Corollary 15.5 says that the shape of a deformed simply-connected plane elastic body is roughly the same as that of the original provided the strain does not attain the critical value. It would be interesting to obtain a higher dimensional analogue of this result.

16. *Complex analysis.* Quasiconformal mappings sometimes arise in function-theoretic problems which appear to be completely unrelated to this class. An excellent example is Teichmüller's theorem [T1] which relates the extremal quasiconformal mappings between two Riemann surfaces with the quadratic differentials on these surfaces.

For a more elementary example, suppose that f is meromorphic in a simply-connected domain D of hyperbolic type in \mathbb{R}^2 and let

$$S_f = \left(\frac{f''}{f'} \right)' - \frac{1}{2} \left(\frac{f''}{f'} \right)^2. \quad (16.1)$$

By a theorem of Nehari [N1], f is injective whenever D is a disk or halfplane and $|S_f| \leq 2\rho_D^2$ in D . Here ρ_D is the hyperbolic metric in D given by

$$\rho_D(z) = |g'(z)|(1 - |g(z)|^2)^{-1} \quad (16.2)$$

where $g: D \rightarrow \mathbb{B}^2$ is conformal. It is natural to ask: for which other domains D does such a result hold? That is, for which D is $\sigma(D) > 0$, where $\sigma(D)$ denotes the supremum of the numbers $a \geq 0$ such that f is injective whenever f is meromorphic with $|S_f| \leq a\rho_D^2$ in D ?

The answer involves quasiconformal mappings and yields a new characterization of Bers's universal Teichmüller space [B5].

16.3. THEOREM. *$\sigma(D) > 0$ if and only if D is a quasidisk.*

17. *Differential geometry and topology.* Some of the results mentioned in Chapter II have important applications in differential geometry. For example, Theorem 1.5 and the necessity in Theorem 10.3 are key steps in the original proof of Mostow's rigidity theorem [M5].

17.1. THEOREM. *If $n > 2$ and if M and M' are diffeomorphic compact Riemannian n -manifolds with constant negative curvature, then M and M' are conformally equivalent.*

Similarly, the equicontinuity property for quasiconformal mappings implied by Theorem 13.1 is an important tool in establishing the following conjecture of Lichnerowicz [L4].

17.2. THEOREM. *If $n \geq 2$ and if M is a compact Riemannian n -manifold not conformally equivalent to a sphere, then the group $C(M)$ of conformal self mappings of M is compact in the topology of uniform convergence.*

The work of Earle and Eells [E1] on the diffeomorphism group of a surface and Bers's proof [B6] of Thurston's theorem on the classification of self mappings of surfaces illustrate how quasiconformal mappings can be applied to problems in surface topology. Sullivan showed [S1] that the Schoenflies theorem, the annulus conjecture and the component problem hold for quasiconformal mappings in all dimensions. The results of this fundamental paper suggest that quasiconformal mappings may prove to be an important, intermediate category of maps between homeomorphisms and diffeomorphisms.

REFERENCES

- [A1] L. V. Ahlfors, *Zur theorie der Überlagerungsflächen*, Acta Math. **65** (1935), 157–194.
- [A2] —, *On quasiconformal mappings*, J. Analyse Math. **3** (1953/54), 1–58.
- [A3] —, *Quasiconformal reflections*, Acta Math. **109** (1963), 291–301.
- [A4] —, *Extension of quasiconformal mappings from two to three dimensions*, Proc. Nat. Acad. Sci. U.S.A. **51** (1964), 768–771.
- [A5] —, *Finitely generated Kleinian groups*, Amer. J. Math. **86** (1964), 413–429.
- [A6] —, *Quasiconformal mappings, Teichmüller spaces, and Kleinian groups*, Proc. Internat. Congr. Math. (Helsinki, 1978), Acad. Sci. Fennica, Helsinki, 1980, pp. 71–84.
- [A7] L. V. Ahlfors and L. Bers, *Riemann's mapping theorem for variable metrics*, Ann. of Math. **72** (1960), 385–404.
- [A8] L. V. Ahlfors and A. Beurling, *Conformal invariants and function-theoretic null sets*, Acta Math. **83** (1950), 101–129.
- [A9] K. Astala, *A remark on quasi-conformal mappings and BMO-functions*, Michigan Math. J. **30** (1983), 209–212.
- [B1] A. Baernstein II and J. J. Manfredi, *Topics in quasiconformal mapping*, Topics in Modern Harmonic Analysis, Istituto Nazionale di Alta Matematica, Roma, 1983, pp. 819–862.
- [B2] L. Bers, *Quasiconformal mappings and Teichmüller's theorem*, Analytic Functions, Princeton Univ. Press, Princeton, N.J., 1960, pp. 89–119.
- [B3] —, *Uniformization by Beltrami equations*, Comm. Pure Appl. Math. **14** (1961), 215–228.
- [B4] —, *Uniformization, moduli, and Kleinian groups*, Bull. London Math. Soc. **4** (1972), 257–300.
- [B5] —, *Quasiconformal mappings, with applications to differential equations, function theory and topology*, Bull. Amer. Math. Soc. **83** (1977), 1083–1100.
- [B6] —, *An extremal problem for quasiconformal mappings and a problem of Thurston*, Acta Math. **141** (1978), 73–98.
- [B7] —, *Finite dimensional Teichmüller spaces and generalizations*, Bull. Amer. Math. Soc. **5** (1981), 131–172.
- [B8] A. Beurling and L. V. Ahlfors, *The boundary correspondence under quasiconformal mappings*, Acta Math. **96** (1956), 125–142.
- [B9] B. Bojarski, *Generalized solutions of a system of first order differential equations of elliptic type with discontinuous coefficients*, Mat. Sb. **43** (1957), 451–503. (Russian)
- [B10] M. Brown, *Locally flat embeddings of topological manifolds*, Ann. Math. **75** (1962), 331–341.
- [C1] L. Carleson, *The extension problem for quasiconformal mappings*, Contributions to Analysis, Academic Press, New York, 1974, pp. 39–47.
- [D1] A. Douady and C. J. Earle, *Conformally natural extension of homeomorphisms of the circle*, Acta Math. (to appear).

- [D2] D. Drasin, *The inverse problem of the Nevanlinna theory*, Acta Math. **138** (1977), 83–151.
- [E1] C. J. Earle and J. Eells, *A fibre bundle description of Teichmüller theory*, J. Differential Geom. **3** (1969), 19–43.
- [E2] C. J. Earle and S. Nag, *Conformally natural reflections in Jordan curves with applications to Teichmüller spaces* (to appear).
- [E3] T. Erkama, *Group actions and extension problems for maps of balls*, Ann. Acad. Sci. Fenn. Ser. A I Math. **556** (1973), 1–31.
- [F1] C. Fefferman, *Characterizations of bounded mean oscillation*, Bull. Amer. Math. Soc. **77** (1971), 587–588.
- [F2] M. H. Freedman and R. Skora, *Strange actions of groups on spheres*, J. Differential Geom. (to appear).
- [F3] B. Fuglede, *Extremal length and functional completion*, Acta Math. **98** (1957), 171–219.
- [G1] D. B. Gauld and J. Väisälä, *Lipschitz and quasiconformal flattening of spheres and cells*, Ann. Acad. Sci. Fenn. Ser. A I Math. **4** (1978/79), 371–382.
- [G2] F. W. Gehring, *Symmetrization of rings in space*, Trans. Amer. Math. Soc. **101** (1961), 499–519.
- [G3] —, *Rings and quasiconformal mappings in space*, Trans. Amer. Math. Soc. **103** (1962), 353–393.
- [G4] —, *Extension of quasiconformal mappings in three space*, J. Analyse Math. **14** (1965), 171–182.
- [G5] —, *Extension theorems for quasiconformal mappings in n -space*, J. Analyse Math. **19** (1967), 149–169.
- [G6] —, *The L^p -integrability of the partial derivatives of a quasiconformal mapping*, Acta Math. **130** (1973), 265–277.
- [G7] —, *Quasiconformal mappings*, Complex Analysis and its Applications. II, International Atomic Energy Agency, Vienna, 1976, pp. 213–268.
- [G8] —, *Injectivity of local quasi-isometries*, Comment. Math. Helv. **57** (1982), 202–220.
- [G9] —, *Extension of quasiisometric embeddings of Jordan curves*, Complex Variables Theory Appl. **5** (1986), 245–263.
- [G10] F. W. Gehring and G. J. Martin, *Discrete quasiconformal groups I*, Proc. London Math. Soc. (to appear).
- [G11] F. W. Gehring and E. Reich, *Area distortion under quasiconformal mappings*, Ann. Acad. Sci. Fenn. Ser. A I Math. **388** (1966), 1–15.
- [G12] F. W. Gehring and J. Väisälä, *The coefficients of quasiconformality of domains in space*, Acta Math. **114** (1965), 1–70.
- [G13] M. Giaquinta, *Multiple integrals in the calculus of variations and nonlinear elliptic systems*, Ann. of Math. Studies, Vol. 105, Princeton Univ. Press, Princeton, N.J., 1983.
- [G14] H. Grötzsch, *Über möglichst konforme Abbildungen von schlichten Bereichen*, Ber. Verh. Sächs. Akad. Wiss. Leipzig **84** (1932), 114–120.
- [I1] T. Iwaniec, *Extremal inequalities in Sobolev spaces and quasiconformal mappings*, Z. Anal. Anwendungen **1** (1982), 1–16.
- [I2] —, *On L^p -integrability in PDE's and quasiregular mappings for large exponents*, Ann. Acad. Sci. Fenn. Ser. A I Math. **7** (1982), 301–322.
- [J1] F. John, *Rotation and strain*, Comm. Pure Appl. Math. **14** (1961), 391–413.
- [J2] —, *On quasi-isometric mappings*, II, Comm. Pure Appl. Math. **22** (1969), 265–278.
- [J3] F. John and L. Nirenberg, *On functions of bounded mean oscillation*, Comm. Pure Appl. Math. **14** (1961), 415–426.
- [J4] P. W. Jones, *Extension theorems for BMO*, Indiana Univ. Math. J. **29** (1980), 41–66.
- [K1] M. Kiiikka, *Diffeomorphic approximation of quasiconformal and quasisymmetric homeomorphisms*, Ann. Acad. Fenn. Sci. Ser. A I Math. **8** (1983), 251–256.
- [K2] I. Kra, *On the Nielsen-Thurston-Bers type of some self-maps of Riemann surfaces*, Acta Math. **146** (1981), 231–270.

- [K3] T. Kuusalo, *Quasiconformal mappings without boundary extensions*, Ann. Acad. Sci. Fenn. Ser. A I Math. **10** (1985), 331–338.
- [L1] M. A. Lavrentieff, *Sur une classe de représentations continues*, Mat. Sb. **42** (1935), 407–423.
- [L2] —, *Sur un critère différentiel des transformations homéomorphes des domaines à trois dimensions*, Dokl. Akad. Nauk. **20** (1938), 241–242.
- [L3] O. Lehto and K. I. Virtanen, *Quasiconformal mappings in the plane*, Springer-Verlag, 1973.
- [L4] J. Lelong-Ferrand, *Transformations conformes et quasi-conformes des variétés riemanniennes compactes (Démonstration de la conjecture de A. Lichnerowicz)*, Acad. Roy. Belg. Cl. Sci. Mém. Collect. **39** (1971), 1–44.
- [L5] —, *Étude d'une classe d'applications liées à des homomorphismes d'algèbres de fonctions, et généralisant les quasi conformes*, Duke Math. J. **40** (1973), 163–186.
- [L6] L. G. Lewis, *Quasiconformal mappings and Royden algebras in space*, Trans. Amer. Math. Soc. **158** (1971), 481–492.
- [L7] C. Loewner, *On the conformal capacity in space*, J. Math. Mech. **8** (1959), 411–414.
- [M1] O. Martio, S. Rickman, and J. Väisälä, *Topological and metric properties of quasi-regular mappings*, Ann. Acad. Sci. Fenn. Ser. A I Math. **488** (1971), 1–31.
- [M2] B. Mazur, *On embeddings of spheres*, Bull. Amer. Math. Soc. **65** (1959), 59–65.
- [M3] D. Menchoff, *Sur une généralisation d'un théorème de M. H. Bohr*, Mat. Sb. **44** (1937), 339–354.
- [M4] C. B. Morrey, *On the solutions of quasi-linear elliptic partial differential equations*, Trans. Amer. Math. Soc. **43** (1938), 126–166.
- [M5] G. D. Mostow, *Quasi-conformal mappings in n -space and the rigidity of hyperbolic space forms*, Inst. Hautes Études Sci. Publ. Math. **34** (1968), 53–104.
- [N1] Z. Nehari, *The Schwarzian derivative and schlicht functions*, Bull. Amer. Math. Soc. **55** (1949), 545–551.
- [N2] R. Nirenberg, *On quasi-pseudoconformality in several complex variables*, Trans. Amer. Math. Soc. **127** (1967), 233–240.
- [R1] H. M. Reimann, *Functions of bounded mean oscillation and quasiconformal mappings*, Comment. Math. Helv. **49** (1974), 260–276.
- [R2] Yu. G. Reshetnyak, *Space mappings with bounded distortion*, Sibirsk. Mat. Zh. **8** (1967), 629–658. (Russian)
- [R3] —, *Liouville's theorem on conformal mappings under minimal regularity assumptions*, Sibirsk. Mat. Zh. **8** (1967), 835–840. (Russian)
- [R4] —, *Stability estimates in Liouville's theorem and the L^p -integrability of the derivatives of quasiconformal mappings*, Sibirsk. Mat. Zh. **17** (1976), 868–896. (Russian)
- [R5] S. Rickman, *Extension over quasiconformally equivalent curves*, Ann. Acad. Sci. Fenn. Ser. A I Math. **436** (1969), 1–12.
- [R6] —, *On the number of omitted values of entire quasiregular mappings*, J. Analyse Math. **37** (1980), 100–117.
- [S1] D. Sullivan, *Hyperbolic geometry and homeomorphisms*, Geometric Topology, Academic Press, New York, 1979, pp. 543–555.
- [S2] —, *On the ergodic theory at infinity of an arbitrary discrete group of hyperbolic motions*, Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference, Ann. of Math. Studies, No. 97, Princeton Univ. Press, Princeton, N.J., 1981, pp. 465–496.
- [S3] —, *Quasiconformal homeomorphisms and dynamics I. Solution of the Fatou-Julia problem on wandering domains*, Ann. of Math. **122** (1985), 401–418.
- [S4] —, *Quasiconformal homeomorphisms and dynamics II: Structural stability implies hyperbolicity for Kleinian groups*, Acta Math. **155** (1985), 243–260.
- [T1] O. Teichmüller, *Extremale quasikonforme Abbildungen und quadratische Differentiale*, Abh. Preuss. Akad. Wiss. Mat.-Nat. Kl. **22** (1940), 1–197.
- [T2] P. Tukia, *On two-dimensional quasiconformal groups*, Ann. Acad. Sci. Fenn. Ser. A I Math. **5** (1980), 73–78.

[T3] —, *A quasiconformal group not isomorphic to a Möbius group*, Ann. Acad. Sci. Fenn. Ser. A I Math. **6** (1981), 149–160.

[T4] —, *Quasiconformal extension of quasimetric mappings compatible with a Möbius group*, Acta Math. **154** (1985), 153–193.

[T5] P. Tukia and J. Väisälä, *Quasiconformal extension from dimension n to $n+1$* , Ann. of Math. **115** (1982), 331–348.

[T6] —, *Bilipschitz extensions of maps having quasiconformal extensions*, Math. Ann. **269** (1984), 561–572.

[V1] J. Väisälä, *On quasiconformal mappings in space*, Ann. Acad. Sci. Fenn. Ser. A I Math. **298** (1961), 1–36.

[V2] —, *On quasiconformal mappings of a ball*, Ann. Acad. Sci. Fenn. Ser. A I Math. **304** (1961), 1–7.

[V3] —, *Lectures on n -dimensional quasiconformal mappings*, Lectures Notes in Math., Vol. 229, Springer-Verlag, 1971.

[V4] —, *A survey of quasiregular maps in \mathbb{R}^n* , Proc. Internat. Congr. Math. (Helsinki, 1978), Acad. Sci. Fennica, Helsinki, 1980, pp. 685–691.

[V5] —, *Quasimöbius maps*, J. Analyse Math. **44** (1984/85), 218–234.

MATHEMATICAL SCIENCES RESEARCH INSTITUTE, BERKELEY, CALIFORNIA 94720, USA

UNIVERSITY OF MICHIGAN, ANN ARBOR, MICHIGAN 48109, USA

Soft and Hard Symplectic Geometry

M. GROMOV

1. Basic definitions, examples, and problems.

1.1. Symplectic forms and manifolds. An exterior differential 2-form ω on a smooth manifold V is called *nonsingular* if the associated homomorphism between the tangent and cotangent bundles of V , denoted $I_\omega: T(V) \rightarrow T^*(V)$ and defined by $I_\omega(\tau) = \omega(\tau, \cdot)$, is an isomorphism. In this case, the dimension of V is necessarily even (we assume that V is finite-dimensional and all connected components of V have the same dimension), say $\dim V = m = 2n$, and the exterior power ω^n (which is a top-dimensional form on V) does not vanish on V . Conversely, if ω^n does not vanish, then ω is nonsingular. For example, every (oriented) area form on a *surface* is nonsingular.

A form ω on V is called *symplectic* if it is nonsingular and closed, that is, $d\omega = 0$. Then (V, ω) is called a *symplectic manifold*.

EXAMPLES. Every surface with an area form is a symplectic manifold. If (V_i, ω_i) are such surfaces for $i = 1, \dots, n$, then the Cartesian product $(V, \omega) = (V_1 \times V_2 \times \dots \times V_n, \omega_1 \oplus \omega_2 \oplus \dots \oplus \omega_n)$ is a $2n$ -dimensional symplectic manifold. The *symplectic area* $\omega(S) = \int_S \omega$ of every surface S in this V equals the sum of the (oriented!) areas of the projections $S \rightarrow V_i$.

An important special case is the *symplectic space* $(\mathbf{R}^{2n}, \omega = \sum_{i=1}^n dx_i \wedge dy_i)$; that is, the sum of n copies of the (x, y) -plane \mathbf{R}^2 with the usual area form $dx \wedge dy$.

A less obvious example is the complex projective space \mathbf{CP}^n which admits a unique (up to a scalar multiple) 2-form ω which is invariant under the action of the unitary group $U(n+1)$ on \mathbf{CP}^n . This form is (easily seen to be) symplectic, and the symplectic area of every surface $S \subset \mathbf{CP}^n$ equals the average number of the intersection points (counted with algebraic multiplicity) of S with hyperplanes $p \subset \mathbf{CP}^n$,

$$\omega(S) = \int_P \#(S \cap p) dp,$$

where P is the (dual) projective space ($\approx \mathbf{CP}^n$) of hyperplanes p in \mathbf{CP}^n and dp is a $U(n+1)$ -invariant measure on P .

1.2. Symplectic (diffeo)morphisms. A C^1 -map $f: (V_1, \omega_1) \rightarrow (V_2, \omega_2)$ is called *symplectic* if the pull-back of ω_2 equals ω_1 ; that is, $f^*(\omega_2) = \omega_1$. Every such f necessarily is an *immersion*; that is, the differential $D_f: T(V_1) \rightarrow T(V_2)$ is injective on the tangent space $T_v(V_1)$ for all $v \in V_1$.

The group $\text{Symp } V$ of symplectic diffeomorphisms of every symplectic manifold $V = (V, \omega)$ is infinite-dimensional for $\dim V \geq 2$. Indeed, the space Ω of closed 2-forms on V essentially is $\{1\text{-forms}/d(\text{functions})\}$ which makes “the functional dimension” of Ω equal $m - 1$ (for $m = \dim V$), while “the functional dimension” of $\text{Diff } V$ is m since every diffeomorphism $V \rightarrow V$ is determined by m functions on V . Hence, the expected “functional dimension” of the group $\text{Symp}(V, \omega)$, which is the isotropy subgroup of $\omega \in \Omega$ for the natural action of $\text{Diff } V$ on Ω , is one. This heuristic argument (which, in fact, can be made precise) suggests some correspondence between functions $V \rightarrow \mathbf{R}$ and symplectic diffeomorphisms of V . (Notice that for “sufficiently nondegenerate” forms ω of degree between 3 and $m - 2$, the automorphism group of (V, ω) is finite-dimensional, and the same is true for *symmetric* differential forms of degree ≥ 2 .)

Symplectic vector fields. A vector field X on (V, ω) is called *symplectic* if the Lie derivative $X\omega$ vanishes. Since

$$X\omega = dI_\omega(X) \quad (1)$$

for all fields X and closed 2-forms ω , the condition $X\omega = 0$ is equivalent to $dI_\omega(X) = 0$. Therefore, for every *closed* 1-form l on V , the field $I_\omega^{-1}(l)$ is symplectic. In particular, for every smooth function (Hamiltonian) $h: V \rightarrow \mathbf{R}$, the field $X = I_\omega^{-1}(dh)$, called a *Hamiltonian field*, is symplectic. *Integrable* Hamiltonian fields X (e.g., those where h has compact support) give us one-parametric subgroups $X_t \subset \text{Symp}(V, \omega)$. This confirms the largeness of $\text{Symp } V$ predicted by the above dimension argument.

Another property of $\text{Symp}(V, \omega)$ which can be seen by looking at Hamiltonians is the *transitivity* of $\text{Symp } V$ on k -tuples of disjoint points in V for every $k = 1, 2, \dots$, where we assume that V is connected. In particular, $\text{Symp } V$ is transitive on V .

COROLLARY (DARBOUX). *Every two symplectic manifolds V_1 and V_2 of the same dimension are locally isomorphic.*

PROOF. If $U_i \subset V_i$, $i = 1, 2$, are sufficiently small neighborhoods, then there obviously exists a *connected* symplectic manifold V' , such that U_i are symplectically diffeomorphic to some neighborhoods $U'_i \subset V'$.

Examples of symplectic diffeomorphisms of $(\mathbf{R}^{2n}, \omega = \sum dx_i \wedge dy_i)$. (1) Every parallel translation of \mathbf{R}^{2n} is symplectic.

(2) Let $z_i = x_i + \sqrt{-1}y_i$ and identify \mathbf{R}^{2n} with \mathbf{C}^n . Then the form ω can be expressed with the Euclidean scalar product by

$$\omega(\tau_1, \tau_2) = \langle \tau_1, \sqrt{-1}\tau_2 \rangle$$

for all (tangent) vectors τ_1 and τ_2 in \mathbf{R}^{2n} . Therefore ω is invariant under unitary transformations of \mathbf{C}^n . In fact, the group $\mathrm{Sp}(2n)$ of *linear* symplectic transformations of \mathbf{R}^{2n} *strictly* contains $U(n)$, as

$$\dim U(n) = n^2 < n(2n + 1) = \dim \mathrm{Sp}(2n).$$

(3) Split \mathbf{R}^{2n} into the sum of n copies of $(\mathbf{R}^2, dx \wedge dy)$ and let f_i for $i = 1, \dots, n$ be area-preserving transformations of \mathbf{R}^2 . Then the Cartesian sum of f_i is symplectic on \mathbf{R}^{2n} .

(4) Identify \mathbf{R}^{2n} with (the total space of) the cotangent bundle of \mathbf{R}^n with coordinates x_1, \dots, x_n . Then the natural action of diffeomorphisms of \mathbf{R}^n on $T^*(\mathbf{R}^n) = \mathbf{R}^{2n}$ is symplectic. Thus $\mathrm{Diff} \mathbf{R}^n$ embeds into $\mathrm{Symp} \mathbf{R}^{2n}$.

One can compose diffeomorphisms in (1)–(4) and also compose them with the above Hamiltonian diffeomorphisms X_t . Thus one obtains symplectic diffeomorphisms of \mathbf{R}^{2n} which may look arbitrarily complicated. Yet, there is no $f \in \mathrm{Symp} \mathbf{R}^{2n}$ sending U_1 into U_2 , where U_1 and U_2 are open subsets in \mathbf{R}^{2n} , such that $\mathrm{Vol} U_1 > \mathrm{Vol} U_2$. In fact, $\mathrm{Vol} U = \int_U \omega^n$ is invariant under $\mathrm{Symp} \mathbf{R}^{2n}$.

PROBLEM. Are there further obstructions besides $\mathrm{Vol} U_1 \leq \mathrm{Vol} U_2$ for the existence of symplectic diffeomorphisms of \mathbf{R}^{2n} mapping U_1 into U_2 ?

It is easy to see that no essentially new obstruction exists for *volume-preserving* diffeomorphisms of \mathbf{R}^{2n} . Namely, if U_1 is relatively compact, the sets $\mathbf{R}^{2n} \setminus U_1$ and U_2 are connected, and

$$\mathrm{Vol} U_1 < \mathrm{Vol} U_2 \quad (*)$$

(the strict inequality removes an irrelevant problem of the boundary behavior of the maps), then there exists a smooth (even real analytic) diffeomorphism f of \mathbf{R}^{2n} sending U_1 into U_2 and satisfying $f^*(\omega^n) = \omega^n$. However, such obstructions do exist for symplectic maps as seen in the following

EXAMPLE. Let U_1 be the round ball in \mathbf{R}^{2n} of radius r , and let U_2 be the ε -neighborhood of a linear subspace $L \subset \mathbf{R}^{2n}$ with $\dim L < n$. Then *there is no symplectic* diffeomorphism (for $\omega = \sum_{i=1}^n dx_i \wedge dy_i$) sending $U_1 \rightarrow U_2$ for $\varepsilon < r$.

The proof (indicated in §4.1) relies on the geometry of *holomorphic curves* for some (nonintegrable) *almost complex* structure in \mathbf{R}^{2n} . No “soft” proof is known at present.

REMARK. If L is the n -dimensional linear subspace in \mathbf{R}^{2n} given by the equations $x_i = 0$, for $i = 1, \dots, n$, then the unit ball U_1 goes to the ε -neighborhood U_2 of L by the symplectic maps $(x_i, y_i) \rightarrow (\varepsilon x_i, \varepsilon^{-1} y_i)$. However, if L is given by the equations $x_1 = 0$ and $y_1 = 0$ (here $\dim L = 2n - 2$), then no symplectic diffeomorphism $U_1 \rightarrow U_2$ exists for $r > \varepsilon$ (see §4.1).

1.3. *Isotropic immersions and Lagrange submanifolds.* A C^1 -map $f: W \rightarrow (V, \omega)$ is called *isotropic* if $f^*(\omega) = 0$. We are especially interested in the case where $\dim W = n$ for $2n = \dim V$ and f is an immersion. These f are called *Lagrange immersions*. Similarly, a submanifold $W \subset V$ is called *Lagrange* if $\omega|_W = 0$.

EXAMPLES. (1) Let $f: (V_1, \omega_1) \rightarrow (V_2, \omega_2)$ be a symplectic map and $(V, \omega) = (V_1 \times V_2, \omega_1 \oplus -\omega_2)$. Then the graph $\Gamma_f: V_1 \rightarrow V$ for $\Gamma_f(v_1) = (v_1, f(v_1))$ is

an isotropic immersion. In fact, the graph of every map is an embedding and $\Gamma_f(V_1)$ is a Lagrange submanifold in V if $\dim V_1 = \dim V_2$.

(2) If $\dim V = 2$ and $W \subset V$ has $\dim W = 1$, then obviously W is Lagrange for every area form ω on V . The Cartesian product of n copies of these,

$$W_1 \times \cdots \times W_n \subset (V_1 \times \cdots \times V_n, \omega_1 \oplus \cdots \oplus \omega_n),$$

also is Lagrange. In particular, the torus $T^n = \{x_i, y_i \mid x_i^2 + y_i^2 = 1\}$ is Lagrange in $(\mathbf{R}^{2n}, \sum_{i=1}^n dx_i \wedge dy_i)$.

(3) Let V be the (total space of the) cotangent bundle, $V = T^*(X)$ for some n -dimensional manifold X . Denote by σ the *canonical* 1-form on V defined by the identity

$$\alpha^*(\sigma) = \alpha, \quad (**)$$

where $\alpha: X \rightarrow T^*(X) = V$ is an arbitrary C^1 -section and where the same α on the right-hand side of $(**)$ is viewed as a 1-form on X . It is easy to see that the form $\omega = d\sigma$ is symplectic and that for $X = \mathbf{R}^n$ this V is symplectically isomorphic to $(\mathbf{R}^{2n}, \sum_{i=1}^n dx_i \wedge dy_i)$. Now, a section $X \rightarrow T^*(X) = V$ is Lagrange if and only if the corresponding 1-form on X is closed. In particular, $dh: X \rightarrow T^*(X)$ is Lagrange for every smooth function h on X . (Notice that every $W \subset T^*(X)$, whose projection on V is a *diffeomorphism* of W onto V , is of this kind; $W = \alpha(X)$ for a unique 1-form $\alpha: X \rightarrow T^*(X) = V$.)

(4) The real projective space $\mathbf{R}P^n \subset \mathbf{C}P^n$ is Lagrange for the above $U(n+1)$ -invariant form ω on $\mathbf{C}P^n$. If $V \subset \mathbf{C}P^n$ is a nonsingular complex algebraic subvariety, then the induced form $\omega' = \omega|_V$ clearly is nonsingular, and hence, symplectic on V . Furthermore, if V is defined over \mathbf{R} , then the *real locus* $W = V \cap \mathbf{R}P^n$ is isotropic in (V, ω') . This W is Lagrange if W is nonsingular and $\dim W = \dim_{\mathbf{C}} V$.

We shall see in §2.2 that the existence problem for Lagrange immersions $\varphi: W \rightarrow V$, for given W and $V = (V, \omega)$, belongs with the soft geometry. But the apparently similar problem of a possible topology of the image $\varphi(W) \subset V$ seems (at the present moment) of hard nature.

EXAMPLE (SEE §3.5). For every Lagrange immersion of a *closed* manifold into $\mathbf{R}^{2n} = \mathbf{C}^n$,

$$\varphi: W \rightarrow \left(\mathbf{R}^{2n}, \omega = \sum_{i=1}^n dx_i \wedge dy_i \right),$$

there exists a nonconstant holomorphic disk in \mathbf{C}^n with boundary in $\varphi(W)$.

It easily follows that the relative cohomology class

$$[\omega] \in H^2(\mathbf{R}^{2n}, W, \mathbf{R})$$

does not vanish. In particular, if $H^1(W; \mathbf{R}) = 0$, then W admits no Lagrange *embedding* into $(\mathbf{R}^{2n}, \omega)$.

2. Symplectic immersions and embeddings.

2.1. Topological obstructions for symplectic immersions. We consider two symplectic manifolds (V, ω) and (W, ω') and ask ourselves when a given continuous map $\varphi: W \rightarrow V$ is homotopic to a symplectic map $f: W \rightarrow V$. There are two obvious obstructions for the existence of f . First, φ must respect the cohomology classes represented by ω and ω' . That is, the homomorphism $\varphi^*: H^2(V; \mathbf{R}) \rightarrow H^2(W; \mathbf{R})$ should send $[\omega]$ to $[\omega']$.

To see the second obstruction we observe that the differential of f , which is a fiber-wise linear map between tangent bundles $D_f: T(W) \rightarrow T(V)$, is *symplectic* in so far as f is symplectic. Here, we call a continuous fiber-wise linear map $\Delta: T(W) \rightarrow T(V)$ symplectic if $\Delta^*(\omega) = \omega'$. (Note that $f^*(\omega) = D_f^*(\omega)$ by the very definition of $f^*(\omega)$.)

Next we consider the space $\{\Delta\}$ of all symplectic fiber-wise linear maps $T(W) \rightarrow T(V)$ and the space $\{\varphi\}$ of continuous maps $W \rightarrow V$. It is easy to see that the projection $\{\Delta\} \rightarrow \{\varphi\}$, which sends each Δ to the underlying φ , is a *Serre fibration*. Hence, the existence of a homotopy between φ and f (where f lifts to $D_f \in \{\Delta\}$) implies the existence of a lift of φ to some $\Delta \in \{\Delta\}$. Note that such a lift is given by a *symplectic homomorphism* of bundles over W , say by

$$\delta: (T(W), \omega') \rightarrow \varphi^*(T(V), \omega)$$

(here $\varphi^*(\)$ denotes the induced bundle and the symplecticity of δ is understood in the obvious sense), and these δ are sections of the fibration over W whose fiber at $w \in W$ consists of linear symplectic maps $T_w(W) \rightarrow T_{\varphi(w)}(V)$. Then we observe, for example, that every φ lifts to some Δ if the manifold W is contractible and $\dim W \leq \dim V$. (This discussion depends only on the nonsingularity of ω and ω' . On the contrary, the earlier condition $\varphi^*[\omega] = [\omega']$ only needs the forms to be closed.)

2.2. IMMERSION THEOREM (SEE [Gr2], [Gr4]). *Let $\varphi: V = (V, \omega) \rightarrow W = (W, \omega')$ be a continuous map which admits a lift to a symplectic map $T(W) \rightarrow T(V)$ and which satisfies $\varphi^*[\omega] = [\omega']$. Then in the following two cases there exists a symplectic map $f: W \rightarrow V$ homotopic to φ .*

(i) OPEN CASE. *The manifold W is open (that is no connected component of W is a closed manifold).*

(ii) EXTRA DIMENSION. $\dim W < \dim V$.

REMARKS. This theorem, obviously, is false if W is closed, V is open, and $\dim W = \dim V$. In fact, not even a topological immersion $W \rightarrow V$ exists in this case.

The map f can be chosen essentially as smooth as the forms ω and ω' . For example, if ω and ω' are C^∞ -smooth (real analytic), then there is some f which is C^∞ (real analytic).

COROLLARY A. *A $2n$ -dimensional manifold (W, ω') admits a symplectic map into $(\mathbf{R}^{2n}, \omega = \sum dx_i \wedge dy_i)$ if and only if the following three conditions are satisfied:*

- (a) *W is open;*
- (b) *the form ω' is exact;*
- (c) *the tangent bundle $(T(W), \omega')$ is a trivial $\mathrm{Sp}(2n)$ -bundle. (This is equivalent to the existence of linearly independent vector fields, say X_i and Y_i on W for $i = 1, \dots, n$, such that $\omega'(X_i, Y_i) = 1$ and $\omega'(X_i, X_j) = \omega'(Y_i, Y_j) = \omega'(X_i, Y_j) = 0$ for $i \neq j$.)*

Note that (a), (b), and (c) are satisfied for every contractible (e.g., homeomorphic to \mathbf{R}^{2n}) manifold W .

COROLLARY B. *A smooth n -dimensional manifold X admits a Lagrange immersion into \mathbf{R}^{2n} if and only if the complexification $T(X) \oplus \sqrt{-1}T(X)$ is a trivial $\mathrm{GL}_n(\mathbf{C})$ -bundle over X .*

Here (as in Corollary A) the “only if” claim is trivial, and “if” follows from Corollary A applied to $W = T^*(X)$.

About the proof of 2.2. First, using a symplectic $\Delta: T(W) \rightarrow T(V)$, one easily constructs a family of local symplectic immersions $f_w: U_w \rightarrow V$ where U_w is a small ball in W around w and f_w is continuous in $w \in W$. Then the required f is assembled out of f_w by appropriately bending (or flexing) these locally defined maps f_w in order to make them agree on the intersections $U_{w_1} \cap U_{w_2}$. The assembling procedure uses (very soft) techniques of topological sheaves (see [Gr4]). It seems unlikely that such an f could be constructed by means of hard analysis.

2.3. Immersions. As we mentioned earlier, every symplectic map $W \rightarrow V$ is an *immersion* but not necessarily an *embedding*, where a map is called an embedding if it is a homeomorphism onto its image. (For example, every immersion without double points is an embedding, provided W is compact.)

We start with a *smooth embedding* $\varphi: W \rightarrow V$ (that is, φ is a smooth immersion as well as an embedding) and try to C^∞ -isotope φ to a *symplectic embedding* $f: W \rightarrow V$. Note that the differential of such an isotopy is a homotopy of fiber-wise linear and fiber-wise injective maps $\Delta_t: T(W) \rightarrow T(V)$, where $\Delta_0 = D_\varphi$ and $\Delta_1 = D_f$ is symplectic.

Now, for a given embedding φ , we *assume* that there exists a homotopy of fiber-wise injective maps $\Delta_t: T(W) \rightarrow T(V)$ (which, for $t > 0$, do not have to be differentials of maps $V \rightarrow W$) such that $\Delta_0 = D$ and Δ_1 is symplectic.

THEOREM A. *If $\varphi^*[\omega] = [\omega']$, then in the following two cases the embedding φ is isotopic to a symplectic embedding $f: W \rightarrow V$.*

- (i) *The manifold W is open and $\dim W < \dim V$.*
- (ii) *$\dim W \leq \dim V - 4$.*

Note that (i) and (ii) exclude the following two cases allowed by the immersion theorem:

- (a) W is open and $\dim W = \dim V$.
- (b) W is closed and $\dim W = \dim V + 2$.

COROLLARY B. *If W is contractible and $\dim W < \dim V$, then every embedding $W \rightarrow V$ is isotopic to a symplectic one. In particular, if W is diffeomorphic to \mathbf{R}^{2n-2} , then there exists a symplectic embedding $(W, \omega') \rightarrow (\mathbf{R}^{2n}, \omega = \sum dx_i dy_i)$.*

REMARK. The above embedding theorem holds true for *proper* symplectic embeddings, provided $\dim W \leq \dim V - 4$. This allows a *proper* symplectic embedding $(W, \omega') \rightarrow (\mathbf{R}^{2n}, \omega = \sum dx_i \wedge dy_i)$ of every W homeomorphic to \mathbf{R}^{2n-4} .

About the proof of Theorem A. Symplectic embeddings $W \rightarrow V$ are obtained by incorporating a (soft) geometric construction of Nash [N1] (used for isometric C^1 -immersions of Riemannian manifolds) into the framework of topological sheaves (see [Gr4]).

About (i) and (ii). The restrictions (i) and (ii) in Theorem A cannot be dropped. This is seen with (hard) analysis of holomorphic curves in V and W (see §4.2).

3. Holomorphic curves in almost complex manifolds.

3.1. Recall that each complex linear structure on \mathbf{R}^{2n} is determined by an automorphism J on \mathbf{R}^{2n} , such that $J^2 = -\text{Id}$, which corresponds to the multiplications by $\sqrt{-1}$. Next, an *almost complex structure* on a manifold V is given by an automorphism of $T(V)$, denoted by J or by $\sqrt{-1}$, such that $J^2 = -\text{Id}$. Almost complex manifolds (V, J) with $\dim V = 2$ are called *Riemann surfaces*.

A C^1 -map between two almost complex manifolds, say $f: (V_1, J_1) \rightarrow (V_2, J_2)$, is called *holomorphic* if the differential D_f is a *complex* linear map between the fibers $T_v(V_1) \rightarrow T_{f(v)}(V_2)$ for all $v \in V$. This is equivalent to the identity $D_f \circ J_1 = J_2 \circ D_f$.

If $\dim V_1 > 2$, then there is no nonconstant holomorphic map $V_1 \rightarrow V_2$ for generic J_1 and J_2 . But if V_1 is a Riemann surface, then, at least locally, there are, roughly, as many holomorphic maps $V_1 \rightarrow V_2$ as in the case $V_1 = \mathbf{C}$ and $V_2 = \mathbf{C}^n$ for $2n = \dim V$. In fact, the equation $D_f \circ J_1 = J_2 \circ D_f$ is *elliptic* in this case. That is, the linearization of this equation is elliptic with the (principal) symbol at every point isomorphic to that of the Cauchy-Riemann equation $\bar{\partial}f = 0$ for maps $\mathbf{C} \rightarrow \mathbf{C}^n$.

DEFINITION. A (parametrized) *holomorphic curve* in an almost complex manifold V is a holomorphic map of a Riemann surface into V , say $f: S \rightarrow V$. Sometimes, we forget the parametrization and deal with *nonparametrized holomorphic curves* $f(S) \subset V$.

3.2. Tame manifolds. We say that a *closed* 2-form ω on V *tames* an almost complex structure J on V if ω is *J-positive*. That is,

$$\omega(\tau, J\tau) > 0 \quad (*)$$

for all nonzero tangent vectors $\tau \in T(V)$. Since every *J*-positive form (obviously) is nonsingular, this ω is symplectic.

EXAMPLES. (A) *Calibrating forms and Kähler manifolds.* Let g be a Riemannian metric on V invariant under J . Then the 2-form

$$\omega = \omega(\tau_1, \tau_2) \stackrel{\text{def}}{=} -g(\tau_1, J\tau_2)$$

satisfies $\omega(\tau, J\tau) = g(\tau, \tau)$. Hence ω is *J*-positive. Such an ω is called *calibrating* (compare [H-L]) if it is closed

If the structure J is *complex* (i.e., integrable) then calibrating forms also are called *Kähler*. For example, the induced form ω' (compare Example 4 in §1.3) on every complex submanifold in $\mathbb{C}P^n$ (obviously) is Kähler.

(B) *Convex functions.* A smooth function $h: V \rightarrow \mathbf{R}$ is called *strictly J-convex* (or plurisubharmonic) if the restriction of h to every holomorphic curve in (V, J) is subharmonic. It is obvious that *J*-convexity of h is equivalent to positivity of the exact 2-form $\omega = dJdh$ on V and that a sufficiently small neighborhood U of each point $v \in V$ admits a strictly *J*-convex function $U \rightarrow \mathbf{R}$. Thus, every (V, J) can be locally tamed by some ω . (But the existence of a local *calibrating* form imposes a nontrivial partial differential condition on J .)

REMARK. Differential forms (of any degree) taming partial differential equations provide a major (if not the only) source of integro-differential inequalities needed for a priori estimates and vanishing theorems. These forms are defined on spaces of jets (of solutions of equations) and they are often (e.g., in Bochner-Weitzenböck formulas) exact and invariant under pertinent (infinitesimal) symmetry groups. Similarly, convex (in an appropriate sense) functions on spaces of jets are responsible for the maximum principles. A great part of hard analysis of P.D.E. will become redundant when the algebraic and geometric structure of taming forms and corresponding convex functions is clarified. (From the P.D.E. point of view, symplectic geometry appears as a taming device on the space of 0-jets of solutions of the Cauchy-Riemann equation.)

3.3. Closed holomorphic curves. If S is a closed Riemann surface, then the space Σ of holomorphic maps $f: S \rightarrow V = (V, J)$ is locally finite-dimensional, since the (elliptic!) Cauchy-Riemann operator is Fredholm. In general, the space Σ is far from compact (even if V is compact), but it admits a nice compactification provided V is a closed manifold (which can be tamed by some closed 2-form ω). This follows (see [Gr3]) from the obvious inequality

$$\text{Area } f(S) \leq \text{const} \cdot \int_S f^*(\omega), \quad (*)$$

where the area is measured with a fixed Riemannian metric on V . Notice that the right-hand side of $(*)$ only depends on the homology class $[f(S)] \in H_2(V)$

and that an inequality similar to (*) holds true for the graph

$$\Gamma_f: S \rightarrow V \times S.$$

EXAMPLES. (A) Let V and S equal the Riemann sphere S^2 . Then the space Σ_1 of holomorphic maps $f: S^2 \rightarrow S^2$ of degree one is noncompact. (This is the group $\mathrm{PGL}_2\mathbf{C}$.) Now we look at (the images of) the graphs of these maps which are smooth holomorphic curves $S_f \subset S^2 \times S^2$. If a sequence of maps $f_i \in \Sigma_1$ diverges, then there is a subsequence, say f_j , such that the curves $S_{f_j} \subset S^2 \times S^2$ converge (in an obvious sense) to a reducible curve in $S^2 \times S^2$ of the form $(S^2 \times s_2) \cup (s_1 \times S^2)$ for some points s_1 and s_2 in S^2 .

(A') Let S be any Riemann surface and V an arbitrary projective algebraic variety. Then the space Σ of holomorphic maps $f: S \rightarrow V$ in a fixed homology class is a quasiprojective variety which can be completed to a projective variety by adding to Σ (graphs of) reducible curves which are obtained by pinching some circles (vanishing cycles) in S .

(B) Let $V = (\mathrm{SL}_2\mathbf{C})/\Lambda$ for some cocompact lattice Λ in the group $\mathrm{SL}_2\mathbf{C}$. Take some nontorsion element $\lambda \in \Lambda$, and let $S = C_\lambda/Z_\lambda$ where $C_\lambda \subset \mathrm{SL}_2\mathbf{C}$ is the centralizer of λ (notice that $C_\lambda \approx \mathbf{C}^*$) and Z_λ is the infinite cyclic group generated by λ . Clearly, this S is a torus which is holomorphically mapped into V and the area of this torus can be made arbitrarily large by applying the (holomorphic) action of $\mathrm{SL}_2\mathbf{C}$ on V . This happens because the complex structure on V is not tame.

SCHWARZ LEMMA. *The bound on area of f provided by (*) allows one, in principle, to control the pointwise norm of the differential Df . For example, every holomorphic map f of the unit disk $B \rightarrow \mathbf{C}$ satisfies*

$$\left| \frac{df}{dz}(0) \right|^2 \leq \pi^{-1} \int_B f^*(\omega) \quad (**)$$

for the area form $\omega = dx \wedge dy$ on \mathbf{C} . (In fact,

$$\left| \frac{df}{dz}(0) \right|^2 \leq \pi^{-1} A,$$

for the area A of the minimal simply connected subset in \mathbf{C} containing the image $f(B) \subset \mathbf{C}$.)

The inequality (**) generalizes to all manifolds (V, J) tamed by *exact* forms and implies (see [Gr3]) that the space of closed holomorphic curves S in every closed tame manifold (V, J) can be compactified by adding (graphs of) singular curves obtained by pinching circles in S .

REMARKS. (a) The pinching of circles in *minimal surfaces* was discovered by Sacks and Uhlenbeck [B-U] (compare [Sch-Y]).

(b) It seems that minimal surfaces become truly useful (for “soft purposes”) in the presence of higher order taming (curvature) forms. Here are two examples.

(b₁) (Frankel Conjecture). *Every closed Kähler manifold V with positive bi-sectional curvature is biholomorphic to \mathbf{CP}^n .*

This is proven by Siu and Yau who start with an appropriate minimal surface S in V . Then they show S is holomorphic and admits as many holomorphic deformations as $CP^1 \subset CP^n$. (This, probably, generalizes to calibrated almost complex manifolds.)

Notice that the algebraic version (due to Hartshorne) of the Frankel conjecture (where “positive curvature” is replaced by “ample tangent bundle” and which is a soft proposition by algebra-geometric standards) was proven by Mori who used, as a hard tool, the action of Frobenius on curves in V (after reducing the problem to finite characteristic).

(b₂) Let V be a Riemannian manifold with positive curvature operator. It is shown in [Mo] and [Mi] that every minimal sphere in V has (Morse) index $\geq m/2 - 3/2$ for $m = \dim V$ and that this inequality implies the homotopy equivalence of the universal covering of V to the sphere S^m , provided that V is closed.

(c) Compactification theorems are known (Uhlenbeck) for many higher-dimensional conformally invariant elliptic systems.

(c') The most fascinating (soft) application of such a compactification (for the Yang-Mills equation, where the hard part had been furnished by Uhlenbeck and Taubes) was discovered by Donaldson [D1]. Also, see [D2, D3, D4, F-S, T1, T2]. (Notice that the Yang-Mills equation on a given principal bundle P over a 4-manifold is tamed by the universal Pontryagin 4-form on the space of jets of connections on P .)

3.4. *Compactness and existence theorems for closed holomorphic curves.* If one rules out the pinching of circles described in the previous section, §3.3, then one concludes to *compactness* of a pertinent space Σ of holomorphic curves.

3.4. A. EXAMPLE. Let $S = S^2$ and consider holomorphic maps $f: S \rightarrow (V, J)$, such that the homology class $f_*[S] \in H_2(V)$ generates the image of the Hurewicz homomorphism $\pi_2(V) \rightarrow H_2(V)$. Then, assuming V is tamed by some ω , no pinching of curves in S is possible (as it would decompose $S \subset V$ into smaller holomorphic spheres). Hence, the space Σ is compact modulo conformal transformations of $S = S^2$. In other words, the space Σ' of corresponding nonparametrized holomorphic curves in V is compact for compact manifolds V . It easily follows (by Fredholm theory for nonlinear elliptic operators) that Σ' represents certain homology class, denoted $[\Sigma']$ in the space of all surfaces in V (here, V is a closed manifold), and that $[\Sigma']$ is invariant under homotopies J_t of J in-so-far as the homotoped structure J_t is tame (i.e., tamed by some ω_t) for all t . Since the space of the almost complex structures tamed by a fixed ω is contractible (this is trivial, see [Gr3]), *the class $[\Sigma']$ is an invariant of the symplectic form.*

Further invariants of (V, ω) are obtained by taking (the homology classes represented by) some subvarieties in Σ' such as the variety Σ'_0 of holomorphic curves $S \subset V$ passing through a fixed point $v_0 \in V$.

REMARK. Our $[\Sigma']$ and $[\Sigma'_0]$ are similar to Donaldson's invariants in gauge theory, but no direct link between the two types of invariants has been found so far. (Compare [Hi].)

3.4. B. Let us compute $[\Sigma'_0]$ in the simplest case, where

$$(V, J) = (V_0 \times S^2, J_0 \oplus J')$$

for a closed aspherical manifold (V_0, J_0) tamed by some form ω_0 , and where (S^2, J') is the standard (Riemann) sphere. For any point $(v_0, s_0) \in V$ there exists a *unique* holomorphic sphere S_0 in V passing through this point and homologous to the sphere $v_0 \times S^2 \in V$. (In fact, $S_0 = v_0 \times S^2$.) It easily follows that the corresponding zero-dimensional class $[\Sigma'_0]$ is nontrivial. Hence, this class also is nontrivial for *every* almost complex structure J_1 on V tamed by the form $\omega = \omega_0 \oplus \omega'$ on V where ω_0 tames J_0 on V_0 and ω' is the standard area form on S^2 .

3.4. B₁. COROLLARY. *For every almost complex structure J_1 on V tamed by ω there exists a holomorphic map $f: S^2 \rightarrow V$ which passes through a given point in V and which is homologous to the sphere $v_0 \times S^2 \subset V$.*

3.4. B₂. REMARK. This corollary is similar to the *Riemann mapping theorem for spheres* which states (in our language) that for every two almost complex structures J_0 and J' on S^2 , there exists a holomorphic sphere

$$S \subset (S^2 \times S^2, J_0 \oplus J')$$

which is homologous to the diagonal in $S^2 \times S^2$ and which passes through three given points in $S^2 \times S^2$, provided that these points are not contained in $(s_1 \times S^2) \cup (S^2 \times s_2) \subset S^2 \times S^2$ for any pair $(s_1, s_2) \in S^2 \times S^2$. In fact, our proof of 3.4.B₁ delivers such a holomorphic S for *every* (nonsplit) structure J_1 on $S^2 \times S^2$ tamed by the form $\omega_0 \oplus \omega'$ on $S^2 \times S^2$, provided $\int_{S^2} \omega_0 = \int_{S^2} \omega'$ and assuming the three points in question are not contained in the union of any two holomorphic spheres S_1 and S_2 in $S^2 \times S^2$ homologous to $s_1 \times S^2$ and $S^2 \times s_2$. (The proof of 3.4.B₁ also yields holomorphic S_1 and S_2 through every point in $(S^2 \times S^2, J_1)$.) This generalization of the Riemann mapping theorem is similar to that due to Schapiro [Sch] and Laurentiev [L]. However, the results in [Sch] and [L] (which are quite general by “hard” standards) are stated and proven in a noninvariant form. (Geometrically speaking, it is assumed in [Sch] and [L] that the spheres $s_1 \times S^2$ and $S^2 \times s_2$ are J_1 -holomorphic for all $(s_1, s_2) \in S^2 \times S^2$.) Besides enormously complicating statements and proofs, this makes these results unsuitable for “soft” applications, though (due to the existence of holomorphic spheres S_1 and S_2 in $(S^2 \times S^2, J_1)$ through every point (s_1, s_2) in $S^2 \times S^2$) the Riemann mapping theorem in [Sch] and [L] is essentially as general as our invariant version. (It seems that hard analytic theorems reach maturity and become truly interesting only when they are applied to a sufficiently large natural class of objects defined in a soft coordinate free language.)

3.4. C. *Solution of nonhomogeneous Cauchy-Riemann equation.* Consider the bundles $H = \text{Hom}(T(S), T(V))$ and $H' = \text{Hom}_{\mathbb{C}}(T(S), T(V))$ over $S \times V$ for

given (almost) complex structures J' in $T(S)$ and J in $T(V)$ and let $h \rightarrow \bar{h}$ denote the quotient homomorphism $H \rightarrow \bar{H} = H/H'$. Then for every smooth section $\varphi: S \times V \rightarrow \bar{H}$ we consider the equation

$$\bar{D}_f = \varphi = \varphi(s, f(s)) \quad (*)$$

for smooth maps $f: S \rightarrow V$.

The graphs $S \rightarrow S \times V$ of solutions f of $(*)$ are holomorphic for some (naturally constructed) almost complex structure $J = J(J', J,)$ on $S \times V$. Hence, the earlier compactness and existence discussion extends to solutions of $(*)$. Furthermore, one shows (see [Gr3]) that the pinching of circles in $S \rightarrow S \times V$ necessarily creates holomorphic spheres in (V, J) . This leads to the following

ALTERNATIVE. *If (V, J) is a closed tame manifold and $S = S^2$ then either there is a nonconstant holomorphic map $S \rightarrow V$ or the equation $(*)$ admits a homotopic to zero solution $f: S \rightarrow V$ for all φ . For example, $(*)$ is always solvable if V is aspherical.*

3.5. Holomorphic curves with boundaries. A submanifold $W \subset (V, J)$ is called *totally real* if $\dim W = \frac{1}{2} \dim V$ and the intersection $T_w(W) \cap JT_w(W) \subset T_w(V)$ equals zero for all $w \in W$. For example, every curve in a Riemann surface is totally real. We say that ω *tames* (V, W) if it tames (V, J) and $\omega|_W = 0$. That is, W is Lagrange (see §1.3) in (V, ω) .

Let S be a compact Riemann surface with boundary and look at holomorphic maps $S \rightarrow V$ sending ∂S to W . The structure of these maps $f: (S, \partial S) \rightarrow (V, W)$ is essentially the same as for the case of a closed surface S . Namely, this space is compactified by singular holomorphic curves obtained by pinching some circles and some arcs in S with boundary points in ∂S . To see the picture, consider a complex manifold V with an \mathbf{R} -structure given by an antiholomorphic involution of V whose fixed point set (the real locus) $W \subset V$ has $\dim W = \frac{1}{2} \dim V$ and, hence, is totally real. Then we take the upper hemisphere $S \subset S^2$ and observe that every holomorphic map $(S, \partial S) \rightarrow (V, W)$ extends, by symmetry, to a (unique) holomorphic map $S^2 \rightarrow V$ commuting with the \mathbf{R} -structures in V and S^2 (where the \mathbf{R} -involution on S^2 is the symmetry in the equator). Thus, the space of holomorphic maps $(S, \partial S) \rightarrow (V, W)$ is identified with the real locus of the space of holomorphic maps $S^2 \rightarrow V$ and the compactification of the former equals the real locus of the latter.

The compactness and existence theorems in §3.4 easily generalize to curves with boundaries (see [Gr3]). In particular, the alternative in §3.4.C holds true for a class of Lagrange submanifolds $W \subset \mathbf{C}^n$ which includes all *closed* submanifolds. For $V = \mathbf{C}^n$ the equation $\bar{D}_f = \varphi$ amounts to $\bar{\partial}f = \varphi$, and if φ is a constant map $S \rightarrow \mathbf{C}^n$ the solutions $f: S \rightarrow \mathbf{C}^n$ are harmonic. Therefore, for compact W , no such f exists if the norm $\|\varphi\|$ is sufficiently large. Hence, by our alternative, there exists a nonconstant holomorphic C^∞ -map of the disk into \mathbf{C}^n with the boundary sent to W for every closed Lagrange C^∞ -submanifold $W \subset \mathbf{C}^n$. A similar reasoning applies to *immersed* Lagrange submanifold W in

\mathbf{C}^n and provides holomorphic disks mentioned in §1.3. Notice that holomorphic disks may be nonsmooth at the boundary points reaching double points of the immersed $W \subset \mathbf{C}^n$. But these disks are Hölder if W has normal crossings.

4. Applications of holomorphic curves in symplectic geometry.

4.1. *Embeddings of open manifolds.* Define *symplectic width* of (V, ω) as the lower bound of the numbers $a > 0$, such that for every almost complex structure J on V tamed by ω and for every point $v \in V$ there exists a nonconstant connected properly mapped J -holomorphic curve $f: S \rightarrow V$ passing through v , such that the symplectic area of f satisfies $\int_S f^*(\omega) \leq a$.

Obviously (see [Gr3]), this width is monotone under *equidimensional* symplectic embeddings. Namely, if V_1 embeds into V_2 , then

$$\text{width } V_1 \leq \text{width } V_2.$$

EXAMPLES. (a) Let U_r be the ball of radius r in \mathbf{C}^n . Every holomorphic curve in U_r through the center has area $\geq r^2$ (this is well known and easy to prove). Hence,

$$\text{width}(U_r, \omega) = \sum dx_i \wedge dy_i \geq \pi r^2.$$

(In fact, $\text{width } U_r = \pi r^2$; see [Gr3].)

(b) Let $(V, \omega) = (V_0 \times S^2, \omega_0 \oplus \omega')$, where (V_0, ω_0) is a $(2n - 2)$ -dimensional closed aspherical manifold and ω' is an area form on S^2 . They by 3.4.B₁,

$$\text{width } V \leq \text{area}(S^2, \omega') \stackrel{\text{def}}{=} \int_{S^2} \omega'.$$

(Clearly, $\text{width } V \geq \text{area } S^2$.)

COROLLARY. *If the above ball U_r embeds into V , then $\text{area } S^2 \geq \pi r^2$.*

Notice that the sphere S^2 minus a point, is symplectically isomorphic to the ε -disk in \mathbf{C} , such that $\pi\varepsilon^2 = \text{area } S^2$ and every relatively compact subset in $\mathbf{C}^{n-1} \times S^2$ embeds into the closed manifold $(\mathbf{C}^{n-1}/\Lambda) \times S^2$ for some lattice Λ in \mathbf{C}^{n-1} . Thus, the corollary shows that U_r symplectically embeds into the ε -neighborhood of the subspace $\{z_1 = 0\} \subset \mathbf{C}^n$ if and only if $r \leq \varepsilon$. This implies the nonembedding results started in §1.2 (compare [Gr3]).

4.2. *Codimension 2 embeddings.* Let V be a symplectic manifold and $V_0 \subset V$ a closed codimension two symplectic submanifold. By using an almost complex structure J on V for which the submanifold V_0 is J -complex, one can show that the invariants Σ' of V (see §3.4) restrict in some natural sense to those of V_0 . A typical corollary one can obtain is as follows:

Let $V = \mathbf{CP}^2 \times \mathbf{R}^2$, where \mathbf{CP}^2 is endowed with the standard (U(3)-invariant) symplectic structure and $\mathbf{R}^2 = (\mathbf{R}^2, dx \wedge dy)$. If a closed 4-dimensional symplectic manifold V_0 symplectically embeds into V , then V_0 is symplectically diffeomorphic to \mathbf{CP}^2 , provided $\pi_2(V_0)$ is cyclic.

4.3. *Existence, extension, and equivalence problems for symplectic structures.* Let ω_0 be a nonsingular 2-form on a connected manifold V .

PROBLEM. Does there exist a homotopy of nonsingular forms, say ω_t for $0 \leq t \leq 1$, such that the form ω_1 is closed (and hence, symplectic)?

If V is open, then the affirmative answer is provided by a sheaf theoretic version of the Smale-Hirsch immersion theory (see [Gr1]). If V is closed, there is an obvious obstruction for the existence of ω_t . Namely, there must exist a 2-dimensional cohomology class α on V , such that $\alpha^n \neq 0$ for $2n = \dim V$. One still does not know if there are further obstructions.

A similar problem is that of extension of a symplectic form from a subset in V to all of V . Here, holomorphic curves provide a nontrivial obstruction.

EXAMPLE (SEE [Gr3]). Let V be an open 4-dimensional manifold such that the Hurewicz homomorphism $\pi_2(V) \rightarrow H_2(V)$ is zero, and let ω_0 be a symplectic form on a neighborhood U of infinity in V . If (U, ω_0) is symplectically diffeomorphic to some neighborhood of infinity in $(\mathbf{R}^4, dx_1 \wedge dy_1 + dx_2 \wedge dy_2)$, and if ω_0 extends to a symplectic form on all of V , then V is diffeomorphic to \mathbf{R}^4 .

Now, consider two symplectic forms ω_0 and ω_1 on a *closed* manifold V which represent the same cohomology class, $[\omega_0] = [\omega_1] \in H^2(V; \mathbf{R})$, and which can be joined by a homotopy of nonsingular forms ω_t . If ω_t is symplectic for all $t \in [0, 1]$, and the cohomology class $[\omega_t]$ is constant in t , then by the Darboux-Moser theorem the manifolds (V, ω_0) and (V, ω_1) are symplectically diffeomorphic. But if $[\omega_t]$ varies, then $[\Sigma']$ -type invariants (see §3.4) may change (see [McD]) and then (V, ω_0) is not symplectically diffeomorphic to (V, ω_1) .

4.4. C^0 -limits of symplectic diffeomorphisms. It was probably assumed in the (“hard minded”) classical mathematics (and mechanics, where symplectic diffeomorphisms are called *canonical transformations*) that symplectic diffeomorphisms can be distinguished from volume-preserving ones by certain global properties stable under uniform limits of diffeomorphisms. This belief (explicitly expressed by Arnold) was confirmed by Eliashberg (see [E2, E3]) who, in particular, proved (by a complicated combinatorial method stemming from Poincaré’s “proof” of his last theorem) that every diffeomorphism of \mathbf{R}^{2n} which is a C^0 -limit of symplectic diffeomorphisms is symplectic. This C^0 -stability also can be derived from the nonembedding results in §4.1 with use of Nash’s implicit function theorem (see [Gr4]).

4.5. Lagrange intersections and fixed points of symplectic diffeomorphisms. As we have seen in §1.2, every symplectic diffeomorphism f lying in a one-parametric subgroup comes from some (generating) function (Hamiltonian) h on the underlying manifold V , provided $H^1(V; \mathbf{R}) = 0$, and the fixed points of f correspond to the critical points of h . Hence, the number of the fixed point of f can be bounded from below by Morse theory. A similar bound (conjectured by Arnold and generalizing the last Poincaré theorem) for exact (which generalizes, in a natural way, the notion of exactness for closed 1-forms corresponding to symplectic fields) area-preserving diffeomorphisms of surfaces was proved by Eliashberg [E1] by his combinatorial method. Then Conley and Zehnder [C-Z] proved another conjecture of Arnold: *every exact symplectic diffeomorphism f of the torus $T^{2n} = \mathbf{R}^{2n}/\mathbf{Z}^{2n}$ has at least $2n + 1$ fixed points, and if f is generic*

then it has at least 4^n fixed points. The proof involves an auxiliary (generating) function L (Lagrangian) on the space of contractible maps $S^1 \rightarrow T^{2n}$ where Conley and Zehnder apply variational techniques similar to those used for locating periodic orbits of Hamiltonian (i.e., symplectic) flows (see [Rab, W2, Ber]). Notice that this L has infinite Morse index (it looks like the quadratic function $\sum_{i=1}^{\infty} x_i y_i$ on C^∞) and the naive Morse theory does not apply to L . (In fact, there are only a few interesting variational problems where the Morse theory based on the Palais-Smale condition can be applied directly.) The method of Conley and Zehnder was cleaned of “hard” analysis by Chaperon who used, instead of L , a function on a finite-dimensional space similar to the space of broken geodesic in a Riemannian manifold. The hard regularity theorem reduces, in Chaperon’s approach, to showing that every broken *extremal* curve is in fact unbroken.

The fixed points of a symplectic diffeomorphism f of (V, ω) correspond to the intersection points of two Lagrange submanifolds in $(V \times V, \omega \oplus -\omega)$, which are the graph of f and the diagonal in $V \times V$. The (Chaperon’s rendition of) Conley-Zehnder method extends to some Lagrange manifolds besides symplectic graphs. For example, Laudenchbach and Sikorav [L-S] established by this method a Morse theoretic lower bound on the number of intersection points of an *exact* (in a suitable sense) Lagrange submanifold in $T^*(X)$ with the zero section $X \subset T^*(X)$, where $T^*(X)$ is endowed with the canonical (see §1.3) symplectic structure. (Compare [Ch, F-W, W1, Z].)

An alternative approach to (self) intersections of immersed Lagrange submanifolds $W \subset V$ is provided by holomorphic curves $(S, \partial S) \rightarrow (V, W)$ (see §3.5 and [Gr3]). Similar curves (in a Morse theoretic framework) are used by Floer [F1] who proved a homological version of a general Arnold’s conjecture. It seems that the method of holomorphic curves has an advantage of greater generality, while the symplectic Morse theory, whenever it applies, leads to finer (Morse) inequalities.

4.6. Contact geometry. A codimension one subbundle $\theta \subset T(x)$ is called *contact* if there exists an almost complex structure J on $X \times \mathbf{R}$, such that

$$\theta = T(X) \cap JT(X)$$

for $X = X \times 0 \subset X \times \mathbf{R}$, and such that X is *strictly J -convex* in $X \times \mathbf{R}$. That is, $X = X \times 0$ is a level of a strictly J -convex function (see §3.2) without critical points. Soft properties of contact manifolds (V, θ) are quite similar to those of symplectic manifolds (see [Gr4]). Fundamental rigidity theorems for contact 3-manifolds were proven by Bennequin [B1] using topological (knot theoretic) techniques. In particular, Bennequin proved the existence of exotic contact structures on \mathbf{R}^3 and on S^3 . (The standard θ on S^3 is $T(S^3) \cap \sqrt{-1}T(S^3)$ for the usual embedding $S^3 \subset \mathbf{C}^2$. The standard \mathbf{R}^3 is obtained by removing a point from this S^3 .)

Now let (V, J) be an almost complex manifold with J -convex boundary Y and let $\theta = T(Y) \cap JT(Y)$. If $W \subset Y$ is totally real in $V \supset Y$, then, under

a suitable taming condition, holomorphic curves in V with boundaries in W behave like those in §3.5. This has a nontrivial effect on the contact geometry of (Y, θ) and, in particular, shows that not every (Y, θ) appears as a tame J -convex boundary of some (V, J) . Using this, one can find, for example, an exotic contact \mathbf{R}^{2n-1} for all $n \geq 1$ (see [Gr3, B2]) and can recapture finer results of Bennequin for 3-manifolds (a private communication by Eliashberg). Yet the holomorphic contact geometry is less understood at the present moment than the symplectic case.

Soft and hard historical remarks. It seems difficult (if possible at all) to assign a precise metamathematical meaning to the notions of relative softness and hardness of an argument or of a theory. Intuitively, “hard” refers to a strong and rigid structure of a given object, while “soft” suggests some weak general property of a vast class of objects. Thus, inequalities and estimates are softer than identities, (algebraic) number theory is harder than analysis (over locally compact fields and adèles), real analysis is softer than complex analysis. Semisimple Lie groups and symmetric spaces look, from a certain angle, almost as hard as the integers, while Riemannian geometry appears, on a whole, nearly as soft as differential topology. The proof of $x^3 + y^3 + z^3 \geq 3xyz$ for $x, y, z \geq 0$ by the convexity of $\log t$ seems softer than the proof by the identity $2(x^3 + y^3 + z^3 - 3xyz) = (x + y + z)((x - y)^2 + (x - z)^2 + (y - z)^2)$, although proofs in algebraic geometry which use elliptic P.D.E. are at a level of hardness with those using Frobenius.

“Soft” and “hard” in this talk are limited to the framework of the *global nonlinear analysis* concerning the geometry of spaces of maps between smooth manifolds. The modern approach to such spaces started with the soft homotopy touch by Serre [S1, S2], and soft techniques and ideas have dominated the theory ever since. It is still not known whether Serre’s results (e.g., the finiteness of the stable homotopy groups of spheres) can be recaptured by hard means, although hard arguments have been occasionally used in similar problems (e.g., Morse theory in Bott periodicity, the action of Frobenius in Adams’ conjecture, linear elliptic operators in the signature theorem and the higher signature conjecture). A similar softness also has prevailed in differential topology, since the flexibility of diffeomorphisms (Thom [Th]), immersions (Smale [Sm]), and surgeries allowed an essential reduction of Diff-problems to the homotopy theory. (This soft topological avalanche has been arrested by Thurston’s geometrization of 3-manifolds and by Donaldson’s gauge invariants of 4-manifolds.)

The softness discovered in topology could not, however, discourage the search for classical hard structures based on nonlinear partial differential equations. However, the naive dream of hard global analysis was destroyed overnight by Nash’s discovery [N1, N2] of the amazing flexibility of solutions of certain nonlinear equations, namely of isometric immersions of Riemannian manifolds. (For example, *every* distance-decreasing embedding of the standard m -sphere into \mathbf{R}^n admits, for $m < n$, a uniform approximation by isometric C^1 -embeddings; see [N1, K].) In fact, one could think (until the work of Donaldson) that Nash’s

phenomenon (which is ubiquitous for nonlinear P.D.E., see [Gr4]) totally rules out any hard P.D.E. structure in the soft vastness of nonlinear function spaces. (The hard structure of *linear* elliptic P.D.E. is firmly rooted today in soft topological soil as had been envisioned by Atiyah [A].) Now, holomorphic curves which are concerned with the most elementary equation (and which logically precede, though historically follow, gauge fields) appear as a first stepping stone to a comprehensive hard nonlinear theory.

REFERENCES

- [A] M. Atiyah, *Global aspects of the theory of elliptic differential operators*, Proc. Internat. Congr. Math. (Moscow, 1966), Izdat. "Mir", Moscow, 1968, pp. 57–65.
- [B1] D. Bennequin, *Entrelacements et équations de Pfaff*, Astérisque **107-108** (1982), 87–163.
- [B2] —, *Problèmes elliptiques, surfaces de Riemann et structures symplectiques*, Sémin. Bourbaki Févr., Soc. Math. de France, 1986, pp. 1–23.
- [Ber] H. Berestycki, *Solutions périodiques de systèmes hamiltoniens*, Sémin. Bourbaki Févr., Soc. Math. de France, 1983.
- [Ch] M. Chaperon, *Quelques questions de géométrie symplectique*, Astérisque **105-106** (1983), 231–249.
- [C-Z] C. Conley and E. Zehnder, *The Birkhoff-Lewis fixed point theorem and a conjecture of V. I. Arnold*, Invent. Math. **73** (1983), 33–49.
- [D1] S. K. Donaldson, *An application of gauge theory to four dimensional topology*, J. Differential Geom. **18** (1983), 279–315.
- [D2] —, *Connection, cohomology and the intersection forms on 4-manifolds*, J. Differential Geom. (to appear).
- [D3] —, *The orientation of Yang-Mills moduli spaces and fundamental groups of 4-manifolds*, Preprint, Oxford Univ.
- [D4] —, *Gauge theory and smooth structures on 4-manifolds*, Preprint, Oxford Univ.
- [E1] Ya. M. Eliashberg, *Estimates on the number of fixed points of area preserving transformation*, Preprint, Syktyvkar, 1978. (Kissian)
- [E2] —, *Rigidity of symplectic and contact structures*, Preprint, 1981.
- [E3] —, *Combinatorial methods in symplectic geometry*, Proc. Internat. Congr. Math. (Berkeley, Calif., USA, 1986).
- [F1] A. Floer (in preparation).
- [F-S] R. Fintushel and R. J. Stern, *SO(3) connections and the topology of 4-manifolds*, J. Differential Geom. **20** (1984), 523–539.
- [F-W] B. Fortune and A. Weinstein, *A symplectic fixed point theorem for complex projective spaces*, Preprint, Univ. of Calif., Berkeley, 1984.
- [Gr1] M. Gromov, *Stable maps of foliations into manifolds*, Izv. Akad. Nauk SSSR Ser. Mat. **33** (1969), no. 4, 707–734.
- [Gr2] —, *A topological technique for the construction of solutions of differential equations and inequalities*, Proc. Internat. Congr. Math. (Nice, 1970), Vol. 2, Gauthier-Villars, Paris, 1971, pp. 221–225.
- [Gr3] —, *Pseudo-holomorphic curves in symplectic manifolds*, Invent. Math. **82** (1985), 307–347.
- [Gr4] —, *Partial differential relations*, Springer-Verlag, Berlin-New York, 1986.
- [H1] N. J. Hitchin, *The self-duality equation on a Riemann surface*, Preprint, Oxford Univ.
- [H-L] R. Harvey and B. Lawson, *Calibrated geometries*, Acta Math. **148** (1982), 48–156.
- [K] N. H. Kuiper, *On C^1 -isometric embeddings. I*, Nederl. Akad. Wetensch. Proc. Ser. A **58** (1955), 545–556.
- [L] M. Lavrentiev, *A fundamental theorem of the theory of quasiconformal mappings of plane regions*, Izv. Akad. Nauk SSSR Ser. Mat. **12** (1948), 513–554.

- [L-S] F. Laudenbach et J.-C. Sikorov, *Persistance d'intersections avec la section nulle au cours d'une isotopie hamiltonienne dans un fibré cotangent*, Invent. Math. **82** (1985), no. 2, 349–358.
- [McD] D. McDuff, *Examples of symplectic structures*, Preprint, SUNY at Stony Brook, 1985.
- [M1] M. J. Micallef, *On the topology of positively curved manifolds*, Preprint, Australian Nat. Univ., 1986.
- [Mo] J. D. Moore, *Minimal two-spheres on the topology of manifolds with positive curvature on totally isotropic two-planes*, Preprint, 1986.
- [N1] J. Nash, *C^1 -isometric embeddings*, Ann. of Math. **60** (1954), no. 3, 383–396.
- [N2] —, *The embedding problem for Riemannian manifolds*, Ann. of Math. **63** (1956), no. 1, 20–63.
- [Rab] P. H. Rabinowitz, *Periodic solutions of Hamiltonian systems: A survey*, MRC Tech. Report n° 2154, Univ. Wisconsin, Madison, 1980.
- [S1] J.-P. Serre, *Homologie singulière des espaces fibrés*, Ann. of Math. **54** (1951), 425–505.
- [S2] —, *Groupes d'homotopie et classes de groupes abéliens*, Ann. of Math. **58** (1953), 258–294.
- [Sch] Z. Schapiro, *Sur l'existence des représentations quasiconformes*, C. R. (Doklady) Acad. Sci. URSS (N.S.) **30** (1941), 690–692.
- [Sm] S. Smale, *The classification of immersions of spheres in Euclidean spaces*, Ann. of Math. (2) **69** (1959), 327–344.
- [S-U] J. Sacks and K. Uhlenbeck, *The existence of minimal immersions of two-spheres*, Ann. of Math. **113** (1981), 1–24.
- [Sch-Y] R. Schoen and S. T. Yau, *Existence of incompressible minimal surfaces and the topology of three dimensional manifolds of non-negative scalar curvature*, Ann. of Math. **110** (1979), 127–142.
- [T1] C. H. Taubes, *Gauge theory on asymptotically periodic 4-manifolds*, Preprint, Harvard University, 1986.
- [T2] —, *Gauge theories and non-linear partial differential equations*, Proc. Internat. Congr. Math. (Berkeley, Calif., USA, 1986).
- [Th] R. Thom, *Quelques propriétés globales des variétés différentiables*, Comment. Math. Helv. **28** (1954), 17–86.
- [W1] A. Weinstein, *Lectures on symplectic manifolds*, Regional Conf. Ser. in Math., No. 29, Amer. Math. Soc., Providence, R.I., 1977.
- [W2] —, *Periodic orbits of convex hamiltonian systems*, Ann. of Math. **108** (1977), 507–518.
- [Z] E. J. Zehnder, *Fixed points of symplectic mappings and periodic solutions of Hamiltonian systems*, Proc. Internat. Congr. Math. (Berkeley, Calif., USA, 1986).

INSTITUT DES HAUTES ÉTUDES SCIENTIFIQUES, 91440 BURES-SUR-YVETTE, FRANCE

Elliptic Curves and Number-Theoretic Algorithms

H. W. LENSTRA, JR.

1. Introduction. In this lecture we shall discuss a problem that has fascinated many mathematicians throughout history, such as Eratosthenes ($\sim -284 - \sim -202$), Fibonacci ($\sim 1180 - \sim 1250$), Fermat (1601–1665), Euler (1707–1783), Legendre (1752–1833), and Gauss (1777–1855). This is the problem of how to find the *prime factor decomposition* of a given large integer.

Surveys of methods that are used for this purpose can be found in Riesel's recent book [27] and in the contributions to [21]. The present lecture is devoted to a development that took place since the appearance of Riesel's book, namely, the introduction of *elliptic curves*.

Two stages can be distinguished in most methods to find the prime factorization of a given number. In the first stage (*primality testing*) one decides whether the number is prime or composite. In the second stage (*factorization*) one finds a nontrivial divisor of the number, if it is composite. It is clear that the complete prime factor decomposition can be obtained by applying a primality testing algorithm and a factorization algorithm recursively. Elliptic curves can be applied both to primality testing and to factorization, and they give rise to algorithms with an excellent performance, both in theory and in practice.

Primality testing is considered to be easier than factorization. Suppose, for example, that two 100-digit numbers p and q have been proved prime; this is easily within reach of the current primality testing methods. Suppose moreover that the numbers p and q are thrown away by mistake, but that the product pq is saved. How to recover p and q ? It must be felt as a defeat for mathematics that, in these circumstances, the most promising approaches are searching the waste paper basket and applying mnemonic-hypnotic techniques.

Until recently, the subject of primality testing and factorization was not taken seriously by most mathematicians. Nowadays, a change in this attitude is noticeable. Partly, this change is due to the introduction of more sophisticated mathematical techniques than were used before. Indeed, the use of elliptic curves, which is the main topic of this lecture, has been referred to as the first application of twentieth-century mathematics to the problem of prime factor decomposition.

Another reason for the increased interest in this area is the possibility to apply number theory to the outside world. The existence and uniqueness of the prime factor decomposition constitute the *fundamental theorem of arithmetic*, and this theorem plays indeed a basic role. For example, a number-theoretical question about a positive integer n —can n be written as the sum of two squares? what is the order of the multiplicative group $(\mathbf{Z}/n\mathbf{Z})^*$?—is considered as settled if it is answered in terms of the prime factorization of n . Given the basic role of the prime factor decomposition in number theory, it seems reasonable to suppose that algorithms to achieve this prime factor decomposition play an important role in possible applications of number theory. To date, the most striking illustration is the cryptographic scheme devised by Rivest, Shamir, and Adleman [28]. For the use of this scheme it is essential that primality testing is *easy*, and the security of the system depends on the fact that factorization is *hard*. It should be remarked that, to a certain extent, this is *negative* application: if a better factoring method is discovered, then the application may cease to exist. This remark should serve as a stimulus for those mathematicians to whom the possibility of applying number theory to the outside world does not appeal and who wish to restore the purity of their science.

To test a given integer $n > 1$ for primality, one usually subjects it to a series of *pseudoprime tests*. Most of these tests are based on a variant of *Fermat's theorem*. This theorem asserts that if n is prime then $a^n \equiv a \pmod n$ for all integers a . These pseudoprime tests have the property that any prime number passes them, but that a composite number is very unlikely to pass them. Hence a single test that n fails to pass suffices to prove that n is composite, although it does not readily yield a factor of n . If, on the other hand, n passes many pseudoprime tests, then it is very likely that n is a prime number. The problem then becomes how to *prove* that n is a prime number. It may be said that the real difficulty of primality testing algorithms is not to *obtain* the answer, “prime” or “composite,” but to prove the *correctness* of the answer, in the case it is “prime.” For this reason one sometimes speaks about primality *proving* algorithms.

If a primality test decides that a number is *not* prime then, as we just noted, it usually does not exhibit a factor of the number. To obtain a factor one applies a factorization algorithm. In contrast to primality testing, the difficulty of factorization is to *obtain* the answer, i.e., a nontrivial divisor of the number; checking the correctness of the answer, once it is obtained, is completely trivial. The total freedom one has in the choice of the method by which to obtain a nontrivial divisor seems to be one of the reasons that there is much more variety in factorization algorithms than in primality tests. Indeed, it is not a priori clear why methods that depend on a mathematical theory would be better than nonmathematical methods, and why factorization should be beyond the abilities of competent clairvoyants or religious officers.

The elliptic curve methods that form the subject of this lecture are best understood as analogue of certain older algorithms, which are discussed in §2. These older algorithms depend on properties of the *multiplicative group*, in particular

on the fact that for a prime number p the order of the multiplicative group $(\mathbf{Z}/p\mathbf{Z})^*$ equals $p - 1$. We remark that the algorithms discussed in §2 are by no means the best algorithms that were used before elliptic curves were introduced; we only discuss them because they are helpful in motivating and understanding the new methods.

Section 3 contains the basic properties of elliptic curves that we need. The best reference is Silverman's recent textbook [35]. As most of the literature on the subject, this book restricts itself to elliptic curves that are defined over *fields*. For our purposes it is more natural, both from a conceptual and from an expository point of view, to work with elliptic curves that are defined over *rings*. The general theory of elliptic curves over commutative rings with 1 can be found in [16, Chapter 2]. In §3 we give the basic definitions, but only in the case that the ring in question satisfies a certain condition; this condition is satisfied, for example, if the ring is a field, and also if the ring is *finite*, which is the case in our applications. This condition allows us to give a very straightforward definition: an elliptic curve is defined by a ternary homogeneous cubic polynomial of a certain normal form; to keep this normal form as simple as possible we assume that 6 is a unit of the ring. The set of points of the curve over the ring is then defined as the set of zeros of this polynomial in a suitably defined projective plane. It is a basic property of elliptic curves that this set of points has the structure of an *abelian group*. It should be remarked that in principle it is possible, by more or less artificial considerations, to avoid elliptic curves over rings that are not fields in the description and analysis of the algorithms that we shall discuss. This was, in fact, done in the original publications [30, 20, 14].

We mentioned above that a number of older primality testing and factorization methods depend on the fact that the order of the multiplicative group $(\mathbf{Z}/p\mathbf{Z})^*$ modulo a prime number p equals $p - 1$. Likewise, in the elliptic curve methods an important role is played by the order of the group $E(\mathbf{Z}/p\mathbf{Z})$ of points of an elliptic curve E over $\mathbf{Z}/p\mathbf{Z}$, for a prime number p . By a theorem of Hasse from 1934, this order is of the form $p + 1 - t$, where t is an integer depending on E and p for which $|t| \leq 2\sqrt{p}$. It may be said that the success of the new methods is due to the fact that, for fixed p , this number t varies if one varies the elliptic curve E . In §4 we discuss several methods to calculate the number t .

In §5 it is explained how to do primality testing with the help of elliptic curves. In particular, we discuss the algorithms of Goldwasser-Kilian [14] and Atkin [2]. Atkin's method is of great practical value, and on most numbers on which it has been tried it is much faster than the previous champion, which is the Cohen-Lenstra version of the test of Adleman, Pomerance, and Rumely [1, 9, 10].

Section 6, finally, describes the elliptic curve factorization method [20]. It is, at the moment, the undisputed champion among factoring methods for the great majority of numbers. The quadratic sieve algorithm of Pomerance [26], which was the previous champion, still seems to perform better on numbers that are built up from two primes of the same order of magnitude. The elliptic curve

method has the very attractive property that its speed depends on the size of the smallest prime divisor of the number n that is being factored: smaller prime factors are easier to find. The quadratic sieve and many other fast factoring algorithms do not have this property; they have a running time that only depends on the size of n and not on the size of its prime factors.

By \mathbf{F}_q we shall denote a finite field of cardinality q . Rings are supposed to be commutative with a unit element, and the latter is supposed to be preserved by ring homomorphisms. The group of units of a ring R is denoted by R^* .

2. Multiplicative methods. In this section we discuss two older algorithms for primality testing and factorization, which depend on properties of the multiplicative group. In practice, these algorithms are not feasible for all numbers, but only if certain conditions are satisfied.

We begin with primality testing. The following theorem is due to Pocklington [24].

THEOREM 1. *Let n be an integer, $n > 1$, and s a positive integer dividing $n - 1$. Suppose that there is an integer a satisfying*

$$a^{n-1} \equiv 1 \pmod{n},$$

$$\gcd(a^{(n-1)/q} - 1, n) = 1 \quad \text{for each prime divisor } q \text{ of } s.$$

Then every prime divisor p of n is $1 \pmod{s}$, and if $s > \sqrt{n} - 1$ then n is prime.

The proof is as follows. Let p be a prime divisor of n , and write $b = (a^{(n-1)/s} \pmod{n})$. From $a^{n-1} \equiv 1 \pmod{n}$ it follows that $b^s \equiv 1 \pmod{p}$, so the order of $(b \pmod{p})$ in the group \mathbf{F}_p^* divides s . Also, if q is a prime divisor of s , then $b^{s/q}$ is not $1 \pmod{p}$, since by hypothesis $a^{(n-1)/q} - 1$ is not divisible by p . Therefore the order of $(b \pmod{p})$ is not a divisor of s/q , for any prime q dividing s , so this order is equal to s itself. By Lagrange's theorem in group theory it follows that s divides $\#\mathbf{F}_p^* = p - 1$. This proves the first assertion of the theorem. If also $s > \sqrt{n} - 1$ then it follows that $p > \sqrt{n}$, and this can only be true for all primes p dividing n if n is prime. This proves Theorem 1.

The use of Theorem 1 in primality testing is as follows. Let n be an integer > 1 that one believes to be prime, for example because it passes pseudoprime tests as described in [17, p. 379; 27, p. 98]. Denote by s the largest divisor of $n - 1$ that one is able to factor completely into primes, and suppose that $s > \sqrt{n} - 1$. Now pick a random nonzero integer $a \pmod{n}$, and test whether it satisfies the two conditions of Theorem 1. Observe that these conditions are easy to test: the prime divisors q of s are known, the powers $a^{n-1} \pmod{n}$ and $a^{(n-1)/q} \pmod{n}$ can be calculated with $O(\log n)$ multiplications and squarings \pmod{n} , and the greatest common divisors can be calculated by means of the Euclidean algorithm. If all conditions are found to be satisfied then it follows from the theorem that n is indeed prime, as required.

It should be mentioned that if n is prime it should not be difficult to find an element $a \in \mathbf{Z}/n\mathbf{Z}$ satisfying the conditions of the theorem. Clearly, any nonzero

$a \in \mathbf{Z}/n\mathbf{Z}$ must satisfy the first condition, if n is prime. It is easy to show that, for fixed q , the second condition is satisfied with probability $1 - q^{-1}$, if n is a given prime and $a \neq 0$ is drawn at random. The probability that a satisfies the second condition for all q may be somewhat smaller, but in any case it is at least $c_0/\log \log n$ for some positive constant c_0 ; also, it is not difficult to prove a slightly more general version of the theorem, in which a is allowed to depend on q .

The basic shortcoming of the primality test based on Theorem 1 is that it can only prove the primality of prime numbers n for which $n - 1$ has a large divisor that one is able to factor completely. This is the case if $n - 1$ has many small prime factors, which happens, for example, for the Fermat numbers $n = 2^k + 1$. Theorem 1 is also useful if $n - 1$ is the product of a small number and a large prime number q ; in the latter case one can attempt to prove the primality of q recursively.

There is an analogue to Theorem 1 with the multiplicative group replaced by a *twisted* multiplicative group. For example, if p is prime then the group $\mathbf{F}_{p^2}^*/\mathbf{F}_p^*$ is a twisted multiplicative group, and it has order $(p^2 - 1)/(p - 1) = p + 1$. This leads to primality tests that can be used for numbers n for which $n + 1$ has a large completely factored divisor. This is the case, for example, for the Mersenne numbers $n = 2^k - 1$. These tests are classically formulated in terms of *Lucas sequences*.

We refer to [27, 38] for the details of these and other generalizations of Theorem 1, and for a description of the primality tests that are based on a combination of the $(n - 1)$ - and $(n + 1)$ -methods. If n has the property that at least one of $n \pm 1$ can be written as the product of a completely factored number and a prime number q that, recursively, has the same property, then the primality of n can be proved by repeated application of the two methods. This method was developed by Selfridge and Wunderlich [32], and they found empirically that it can be applied to most primes of at most 35 digits, if “completely factored” is taken to mean “built up from primes below 30030.” The generalizations due to Williams et al. [38] can be used for most prime numbers of at most 80 digits.

The advantage of elliptic curves in this context is that there are so many of them. Each elliptic curve gives rise to a group, and the order of this group varies with the curve. Instead of using the numbers $n \pm 1$, one uses essentially a random number in the neighborhood of n , and one can keep changing the curve until this number factors in the desired way. We refer to §5 for more details.

Next we consider a factorization method that also depends on the multiplicative group. It was invented by Pollard [25], and it is known as the *Pollard $(p - 1)$ -method*.

The Pollard $(p - 1)$ -method attempts to find a nontrivial divisor of a composite integer $n > 1$ in the following way. Pick $a \in \mathbf{Z}/n\mathbf{Z}$ at random, and select a positive integer k that is divisible by many small prime powers; for example, one can take $k = \text{lcm}\{1, 2, \dots, w\}$ for a suitable bound w . Next one calculates

$a_k = (a^k \bmod n)$. This can be done by performing $O(\log k)$ squarings and multiplications $(\bmod n)$. Finally, one calculates $\gcd(a_k - 1, n)$ by means of Euclid's algorithm, and one hopes that this \gcd is a nontrivial divisor of n .

Pollard's $(p - 1)$ -method is usually successful if n has a prime divisor p for which $p - 1$ is built up from small prime factors only. Suppose, to be specific, that $p - 1$ divides k , and that p does not divide a . Since the order of $(\mathbf{Z}/p\mathbf{Z})^*$ equals $p - 1$, it then follows that $a^k \equiv 1 \bmod p$, so p divides $\gcd(a_k - 1, n)$. In many cases one has $p = \gcd(a_k - 1, n)$, and the method finds a nontrivial divisor of n .

Along these lines it can be proved that the Pollard $(p - 1)$ -method is good in discovering prime divisors p of n for which $p - 1$ has no large prime factors. It can also be proved that if n has no such prime divisor p then the method is unlikely to work within a reasonable amount of time.

We refer to [25] for a refinement of the method, which improves its practical performance; to [39] for a variant that uses a twisted multiplicative group, and for which $p + 1$ rather than $p - 1$ should be built up from small prime factors; and to [3] for a generalization that appears to be only of theoretical value.

The advantage of elliptic curves is the same as with primality testing. If one uses an elliptic curve rather than the multiplicative group, then $p \pm 1$ is replaced by a number in the neighborhood of p that varies with the curve, and one can keep changing the curve until the algorithm is successful; one may hope that a fair proportion of the numbers in the neighborhood of p is built up from small primes only, so that not too many curves need be tried. More details can be found in §6.

3. Elliptic curves over rings. Let R be a ring. A finite collection $(a_i)_{i \in I}$ of elements of R will be called *primitive* if it generates R as an R -ideal, i.e., if there exist $b_i \in R$, for $i \in I$, such that $\sum_{i \in I} b_i a_i = 1$. This terminology will in particular be applied to *vectors* and to *matrices* that have coefficients in R . Notice that if R is a field, a collection $(a_i)_{i \in I}$ is primitive if and only if not all a_i are zero.

In the sequel we assume that R satisfies the following two conditions:

- (i) $6 \in R^*$;
- (ii) for all positive integers n, m and every primitive matrix $(a_{ij})_{1 \leq i \leq n, 1 \leq j \leq m}$ over R with the property that all 2×2 -subdeterminants vanish ($a_{ij}a_{kl} - a_{il}a_{kj} = 0$ for all i, j, k, l with $1 \leq i < k \leq n, 1 \leq j < l \leq m$) there exists an R -linear combination of the rows that is primitive as an element of R^m .

If R is a field the first condition means that $\text{char } R \neq 2, 3$. We impose this condition only to simplify the exposition; for $6 \notin R^*$ one must work with more general normal forms for elliptic curves, as in [35, Chapter 3].

The second condition, however, is essential for the definition of elliptic curves and their addition law that we shall give. Condition (ii) means that every *projective R -module* of rank one is free, or equivalently that the *Picard group* $\text{Pic } R$ of R vanishes [4]. Obviously, the condition is satisfied for fields, and below

we shall see that it is also satisfied for finite rings. More generally, it holds for rings that have only finitely many maximal ideals. If R is a Dedekind ring, for example, the ring of integers in a number field, then (ii) is true if and only if the class group of R is trivial.

It is easy to prove that the primitive element of R^m whose existence is postulated by (ii) is in fact uniquely determined up to multiplication by units.

Let R be a ring satisfying (i) and (ii). The unit group R^* acts on the set of primitive triples $(x, y, z) \in R^3$ by $u(x, y, z) = (ux, uy, uz)$. The set of orbits under this action is denoted by $\mathbf{P}^2(R)$, and called the *projective plane* over R . The orbit of (x, y, z) is denoted by $(x : y : z)$.

An *elliptic curve* over R is a pair of elements $a, b \in R$ for which $4a^3 + 27b^2 \in R^*$. These elements are to be thought of as the coefficients in the homogeneous Weierstrass equation

$$y^2z = x^3 + axz^2 + bz^3.$$

We denote the elliptic curve (a, b) by $E_{a,b}$, or simply by E . If we multiply the above equation by u^6 , for some $u \in R^*$, and replace u^2x, u^3y by x, y , respectively, then we obtain the equation for $E_{a',b'}$, where $a' = u^4a$ and $b' = u^6b$. Two such curves are said to be *isomorphic* over R .

Let $E = E_{a,b}$ be an elliptic curve over R . The *set of points* $E(R)$ of E over R is defined by

$$E(R) = \{(x : y : z) \in \mathbf{P}^2(R) : y^2z = x^3 + axz^2 + bz^3\}.$$

The point $(0 : 1 : 0) \in E(R)$ is called the *zero point* of the curve, and denoted by O . Notice that if R is a field this is the only element of $E(R)$ whose z -coordinate is zero.

It is a basic fact that $E(R)$ has in a natural way the structure of an *abelian group* with O as the neutral element. The group law, which is written additively, is such that $-(x : y : z) = (x : -y : z)$ for all $(x : y : z) \in E(R)$. To define the group law we first consider the case that R is a *field*. In this case the addition formulae, and the proof that $E(R)$ is a group, can be found in [35, Chapter 3]. We briefly summarize what we need.

Let R be a field, and let $P_1, P_2 \in E(R)$. To add P_1 and P_2 , consider the straight line passing through P_1 and P_2 (the tangent line to the curve if $P_1 = P_2$). The line and the curve have three intersection points, if we count them with suitable multiplicities, and two of them are P_1 and P_2 . If Q is the third one, then $P_1 + P_2 = -Q$. To turn this geometric description into algebraic formulae, we may suppose that P_1 and P_2 are nonzero and that $P_1 \neq -P_2$. Then we can write $P_i = (x_i : y_i : 1)$ for $i = 1, 2$, where (x_i, y_i) lie on the affine curve $y^2 = x^3 + ax + b$. The straight line is given by $y = \lambda x + \nu$, where

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{or} \quad \lambda = \frac{x_2^2 + x_2x_1 + x_1^2 + a}{y_2 + y_1}$$

and $\nu = y_1 - \lambda x_1$. Notice that $P_1 \neq -P_2$ implies that at least one of the values for λ is well defined, and that they are equal if they are both well defined. The

sum $P_3 = P_1 + P_2$ is now given by $P_3 = (x_3 : y_3 : 1)$, where

$$x_3 = \lambda^2 - x_1 - x_2, \quad y_3 = -(\lambda x_3 + \nu).$$

This gives the addition formulae if R is a field, but for the sequel it is desirable to bring them into homogeneous form. To do this, one replaces x_i and y_i by x_i/z_i and y_i/z_i , respectively, and one clears the denominators. Then one finds that the sum of two points $P_1 = (x_1 : y_1 : z_1)$, $P_2 = (x_2 : y_2 : z_2)$ on $E(R)$ is given by one of two formulae $(q_1 : r_1 : s_1)$, $(q_2 : r_2 : s_2)$, depending on which formula for λ is used. Here q_1, \dots, s_2 are certain polynomial expressions in $x_1, y_1, z_1, x_2, y_2, z_2, a$ with integer coefficients. It turns out that for every pair $(P_1, P_2) \in E(R) \times E(R)$ except $(P_1, P_2) = (O, O)$ at least one of these two formulae is meaningful in the sense that it does not give $(0 : 0 : 0)$, and that any of the two that is meaningful actually gives the sum of P_1 and P_2 in the group $E(R)$. For the remaining pair (O, O) we know of course that $O + O = O = (0 : 1 : 0)$, but this formula is not satisfactory because it does *not* have the property of correctly giving the sum $P_1 + P_2$ for all pairs of points P_1, P_2 for which it is meaningful. To remedy this situation one has to develop an addition law that is valid "in a neighborhood of (O, O) ," and that can be done as in [35, Chapter IV, §1]. The result is that one finds nine polynomial expressions q_i, r_i, s_i ($i = 1, 2, 3$) in $x_1, y_1, z_1, x_2, y_2, z_2, a, b$ with integer coefficients, with the property that the sum of *any* two points $P_1 = (x_1 : y_1 : z_1)$, $P_2 = (x_2 : y_2 : z_2)$ on $E(R)$ is given by one of the three formulae $(q_i : r_i : s_i)$, $i = 1, 2, 3$, and that in fact any of the three formulae that is meaningful is correct. The latter statement is equivalent to nine formal identities $q_1 r_2 - q_2 r_1 = 0, \dots, r_2 s_3 - r_3 s_2 = 0$ in the ring $\mathbf{Z}[a, b, x_1, y_1, z_1, x_2, y_2, z_2]/I$, where a, \dots, z_2 are considered as polynomial variables and I denotes the ideal generated by the two polynomials $y_i^2 z_i - x_i^3 - a x_i z_i^2 - b z_i^3$, $i = 1, 2$. Likewise, the fact that $P_1 + P_2$ lies again on the curve, and that the addition defined in this way satisfies the group axioms, with the zero element and the negatives of points as indicated above, is expressed by a series of formal identities in the same ring. Nine explicit polynomials q_1, \dots, s_3 with all these properties can be found in [19].

We now drop the condition that R be a field. To add two points $P_1 = (x_1 : y_1 : z_1)$, $P_2 = (x_2 : y_2 : z_2)$ on $E(R)$ one proceeds as follows. One uses the same nine polynomial expressions that appeared above to obtain a 3×3 -matrix

$$\begin{pmatrix} q_1 & r_1 & s_1 \\ q_2 & r_2 & s_2 \\ q_3 & r_3 & s_3 \end{pmatrix}$$

with entries from R . This is a primitive matrix, since otherwise there would be a maximal ideal $\mathfrak{m} \subset R$ containing all nine entries; but this would contradict the fact that at least one of the rows can be used to add the two points $P_1 \bmod \mathfrak{m}, P_2 \bmod \mathfrak{m}$ on the elliptic curve $E_{a \bmod \mathfrak{m}, b \bmod \mathfrak{m}}(R/\mathfrak{m})$ over the field R/\mathfrak{m} . Also, all 2×2 -subdeterminants of the matrix are zero, so by condition (ii) above there is an R -linear combination (q_0, r_0, s_0) of the rows that is primitive;

moreover, the orbit of (q_0, r_0, s_0) under R^* is uniquely determined. We now define the sum of P_1 and P_2 on $E(R)$ to be $(q_0 : r_0 : s_0)$.

The fact that $E(R)$ is closed under this operation, and that the addition defined in this way satisfies the group axioms, with the zero element and the negatives of points as indicated earlier, is a consequence of the formal identities that we mentioned above. We omit the details, which are somewhat tedious.

It is a natural question to ask for an *algorithm* to add two points on $E(R)$. From the definition of addition we see immediately that, given the formulae from [19], it suffices to have an algorithmic version of condition (ii): one needs a method to *find* the primitive linear combination that is asserted to exist. Before we describe such a method for the case that R is *finite* it should be pointed out that at the moment this method has only theoretical value. Namely, for the purposes that we have in mind (see the following sections) there is a much easier method, as follows. Pick any nonzero entry from the matrix, and determine whether it is a unit in R . If it is, then the row containing that element is primitive, and one is done. If it isn't, then one knows a nonzero nonunit of R , and in each of the cases that we shall consider this is also satisfactory. Suppose for example, that $R = \mathbf{Z}/n\mathbf{Z}$, where n is an integer that one is trying to factor; then a nonzero nonunit of R leads to a nontrivial divisor of n , which is exactly what one wants.

Assume now that R is a *finite* ring. We assume that the elements of R are represented by elements of a certain finite set S ; one may think of S , for example, as consisting of strings of zeros and ones. It is allowed that two distinct elements s, s' of S represent the same element of R , but we do require that given $s, s' \in S$ there is an efficient algorithm to decide whether this is the case. Here "efficient" may be taken to mean that the time needed by the algorithm is bounded by a polynomial function of $\log \#S$. We also require that there is an efficient algorithm to do *addition* in R ; that is, given $s, s' \in S$, one should be able to find an element of S that represents the sum of the elements represented by s and s' . Likewise we require that *subtraction* and *multiplication* can be done efficiently, as well as the solution of equations of the sort $cx = d$ (given c and d , find x), if they are solvable. Finally we require that an element representing $1 \in R$ is known.

With these hypotheses there is an efficient algorithm that given a primitive $n \times m$ -matrix (a_{ij}) as in condition (ii) produces a linear combination of the rows that is primitive; here "efficient" means that the time needed by the algorithm is bounded by a polynomial function of n, m , and $\log \#S$. We begin with a lemma.

LEMMA. *Let R, S be as above, and denote by t the least positive integer for which $2^{t+1} > \#S$. Then for every $c \in R$ there exists $x \in R$ with $c^{t+1}x = c^t$. Moreover, an element $c \in R$ is nilpotent if and only if $c^t = 0$.*

PROOF. Consider the sequence of ideals

$$R \supset Rc \supset Rc^2 \supset \cdots \supset Rc^t \supset Rc^{t+1}.$$

If any two consecutive ideals in this chain are distinct, one obtains $\#S \geq \#R \geq \text{index}[R: Rc^{t+1}] \geq 2^{t+1}$, which is a contradiction. Hence $c^i = c^{i+1}x$ for some $x \in R$ and some integer i with $0 \leq i \leq t$, and the first statement of the lemma follows upon multiplication by c^{t-i} .

If u is an integer with $u > t$, then it follows that $c^u x = c^{u-1}$. Therefore, if c is nilpotent, the smallest integer u with $c^u = 0$ cannot be larger than t . This implies the last statement of the lemma.

It follows from the lemma that there is an efficient algorithm to decide whether an element of the ring is nilpotent.

We now describe an efficient algorithm that given an $n \times m$ -matrix $A = (a_{ij})$ as in (ii) finds a primitive combination of its rows. The algorithm proceeds by recursion on the cardinality of R . If R is the zero ring (which can be decided by testing whether $1 = 0$, where $0 = 1 - 1$), then any row of the matrix is primitive. Now suppose that R is not the zero ring. Since the matrix is primitive, not all of its entries are nilpotent. Let c be an entry that is not nilpotent. Using the lemma, solve $c^{t+1}x = c^t$. Then $c^{2t}x^t = c^t$, so if we put $e = c^t x^t$ then e is an idempotent: $e^2 = e$. Also, from $c^t e = c^t \neq 0$ one sees that $e \neq 0$. If now $e = 1$ then c is a unit, so the row of the matrix containing c is primitive, and one is done. Suppose therefore that $e \neq 1$. Then $R_1 = Re$ and $R_2 = R(1 - e)$ are nonzero commutative rings with unit elements e and $1 - e$, respectively. Moreover, the map $R \rightarrow R_1 \times R_2$ sending $r \in R$ to $(re, r(1 - e))$ is an isomorphism of rings. The matrix A gives rise to a matrix A_1 over R_1 and a matrix A_2 over R_2 . Now notice that, for each $i = 1, 2$, the map $S \rightarrow R \rightarrow R_i$ shows that the set S can again be used to represent the elements of R_i , and that the same conditions as for R are satisfied. Hence, recursively, we can find an R_i -linear combination of the rows of A_i that is primitive as an element of R_i^m , for each $i = 1, 2$. Adding these two rows in R^m one finds the desired primitive linear combination of the rows of A . This finishes the description of the algorithm.

We remark that, in the above algorithm, the element $c \in R$ is mapped to an element $(c_1, c_2) \in R_1 \times R_2$ for which c_1 is a unit and c_2 is nilpotent. Hence the row of A_1 containing c_1 is already primitive, and the recursion is only needed for the ring R_2 . Since the number of nilpotent entries in A_2 is at least one more than in the matrix A , this shows that the depth of the recursion is bounded by nm . In the case that is of interest to us one has $nm = 9$.

4. The number of points on an elliptic curve. Let R be a finite ring with $6 \in R^*$, and $E = E_{a,b}$ an elliptic curve over R . In this section we discuss the order of the finite group $E(R)$.

If $f: R \rightarrow R'$ is any ring homomorphism from R to a ring R' that also satisfies the two conditions (i), (ii) from §3, then $E_{f(a), f(b)}$ is an elliptic curve over R' . We denote this elliptic curve again by E .

If R contains an element c that is neither a unit nor nilpotent then, as we saw in the previous section, R can be written as the product of two nonzero rings. By induction on $\#R$ it follows that R is isomorphic to the product of

finitely many rings R_i , where each R_i is such that every element of R_i is either nilpotent or a unit. Then each R_i is a *local* ring, which means that the set \mathfrak{m}_i of nonunits of R_i forms an ideal of R_i ; this ideal must be maximal, so that R_i/\mathfrak{m}_i is a field. It is now easy to see that $E(R)$ is isomorphic to the product of the groups $E(R_i)$, so that $\#E(R) = \prod_i \#E(R_i)$. Furthermore, from Hensel's lemma one can deduce that for each i the natural group homomorphism $E(R_i) \rightarrow E(R_i/\mathfrak{m}_i)$ is *surjective* and that its kernel has the same cardinality as \mathfrak{m}_i , so that $\#E(R_i) = \#E(R_i/\mathfrak{m}_i) \cdot \#\mathfrak{m}_i$. Summarizing, we have

$$\frac{\#E(R)}{\#R} = \prod_{\mathfrak{m}} \frac{\#E(R/\mathfrak{m})}{\#R/\mathfrak{m}},$$

where \mathfrak{m} ranges over the set of maximal ideals of R . If these maximal ideals are known, then this formula reduces the computation of $\#E(R)$ to the case that R is a field. If $R = \mathbf{Z}/n\mathbf{Z}$ for some positive integer n , then the above formula reads

$$\frac{\#E(\mathbf{Z}/n\mathbf{Z})}{n} = \prod_p \frac{\#E(\mathbf{F}_p)}{p},$$

where p ranges over the set of primes dividing n . Notice that the same formula holds with the order of the elliptic curve replaced by the Euler ϕ -function, which is the order of the multiplicative group.

Assume, for the rest of this section, that R is a finite *field*, of characteristic different from 2 and 3. Denote the cardinality of R by q , so that we may write $R = \mathbf{F}_q$. We assume that an explicit representation for the elements of R is available, as in the previous section, and that each arithmetic operation in R can be performed in time $O((\log q)^2)$.

According to a theorem of Hasse (1934) we have $\#E(\mathbf{F}_q) = q + 1 - t$, where t is an integer satisfying $|t| \leq 2\sqrt{q}$. Four methods have been proposed to calculate the number $\#E(\mathbf{F}_q)$ or, equivalently, the number t .

The first method, which was employed by Lang and Trotter [18], depends on the formula

$$\#E(\mathbf{F}_q) = 1 + \sum_{x \in \mathbf{F}_q} (1 + \chi(x)),$$

where $\chi(x)$ denotes the element of $\{0, 1, -1\}$ that maps to $(x^3 + ax + b)^{(q-1)/2}$ under the natural map $\mathbf{Z} \rightarrow \mathbf{F}_q$. To prove this formula one simply notes that, for fixed $x \in \mathbf{F}_q$, the number of $y \in \mathbf{F}_q$ with $y^2 = x^3 + ax + b$ is given by $1 + \chi(x)$. Applying this formula in a straightforward way leads to an algorithm to calculate $\#E(\mathbf{F}_q)$ that takes time $O(q^{1+\varepsilon})$, for any $\varepsilon > 0$.

The second method, which is significantly faster, is *probabilistic* in the sense that it depends on random choices. It is analogous to an algorithm of Shanks [33] for the calculation of class numbers of imaginary quadratic fields. We give a brief description.

First, one picks a random point $P \in E(\mathbf{F}_q)$. This is done by selecting random elements $x \in \mathbf{F}_q$ until an element is found for which $x^3 + ax + b$ is a square in \mathbf{F}_q ; this can be tested by checking whether $\chi(x) \neq -1$, with χ as above. If such

an x has been found, one can find an element $y \in \mathbf{F}_q$ with $y^2 = x^3 + ax + b$ by applying another probabilistic algorithm of Shanks [34] or by applying a general zero-finding routine for polynomials over finite fields [17, §4.6.2]. The point $P = (x : y : 1)$ is now on the curve.

Next one determines all integers m for which both $|m - (q + 1)| \leq 2\sqrt{q}$ and $m \cdot P = O$. Clearly such integers exist, since $m = \#E(\mathbf{F}_q)$ has these properties. By means of the “baby step–giant step” strategy, for the details of which we refer to [33], all these integers m can be found in time $O(q^{(1/4)+\varepsilon})$, for any $\varepsilon > 0$.

If m is unique, then $m = \#E(\mathbf{F}_q)$, and one is done. If m is not unique, then the difference between any two consecutive m ’s equals the order of P , and it is easy to see that P cannot generate the group $E(\mathbf{F}_q)$, if $q \geq 37$. In the latter case one selects another random point $P' \in E(\mathbf{F}_q)$, and in a similar way one determines the order of the point P' modulo the subgroup generated by P . In this way one continues until the order k of the subgroup that has been found satisfies $|k - (q + 1)| \leq 2\sqrt{q}$. Then $\#E(\mathbf{F}_q) = k$, if $q \geq 37$.

This algorithm has expected running time $O(q^{(1/4)+\varepsilon})$, for any $\varepsilon > 0$, and it determines not only the order of $E(\mathbf{F}_q)$ but also its group structure. It is of practical value if q has not more than approximately 20 decimal digits.

The third method that we discuss is due to Schoof [30]. It is completely deterministic. The method depends on properties of the *Frobenius endomorphism* ϕ of the curve, which is defined as follows. Denote by K an algebraic closure of \mathbf{F}_q . Then ϕ is the automorphism of the abelian group $E(K)$ defined by

$$\phi(x : y : z) = (x^q : y^q : z^q).$$

Notice that $E(\mathbf{F}_q)$ may be considered as a subgroup of $E(K)$, and that $E(\mathbf{F}_q) = \{P \in E(K) : \phi(P) = P\}$. It is a basic theorem that ϕ satisfies the quadratic equation $\phi^2 - t\phi + q = 0$ in the endomorphism ring of $E(K)$, where t is the integer for which $\#E(\mathbf{F}_q) = q + 1 - t$.

To determine t one now observes that it suffices to determine $t \bmod l$ for all odd primes $l \leq c_1 \log q$ that are different from $\text{char } \mathbf{F}_q$; here c_1 is a positive constant, chosen such that $\prod l > 4\sqrt{q}$ for all q . Namely, if one knows all these $t \bmod l$ then one can determine $t \bmod \prod l$ by means of the Chinese remainder theorem, and since $|t| \leq 2\sqrt{q}$ this suffices to find t and hence $\#E(\mathbf{F}_q)$.

Now let l be an odd prime number, $l \neq \text{char } \mathbf{F}_q$. To determine $t \bmod l$, one first calculates the polynomial ψ_l defined by

$$\psi_l = l \cdot \prod (X - x),$$

with x ranging over the set of those elements of K for which there exists $y \in K$ for which $(x : y : 1)$ is an element of $E(K)$ of order l . It is known that ψ_l has degree $(l^2 - 1)/2$ and belongs to $\mathbf{F}_q[X]$. The polynomial ψ_l can be calculated by means of recursion formulae that can be found, for example, in [35, Chapter III, Exercise 3.7].

Define the ring T by

$$T = \mathbf{F}_q[X, Y]/(\psi_l, Y^2 - X^3 - aX - b).$$

Every element of T has a unique representation

$$\sum_{i=0}^{(l^2-3)/2} \sum_{j=0}^1 a_{ij} \overline{X}^i \overline{Y}^j \quad \text{with } a_{ij} \in \mathbf{F}_q,$$

where $\overline{X}, \overline{Y}$ denote the images of X, Y in T . It follows that T is a finite ring in which the ring operations can be performed efficiently, in the sense of §3.

Let $Q = (\overline{X} : \overline{Y} : 1) \in E(T)$, and define the endomorphism $\sigma: E(T) \rightarrow E(T)$ by the same formula as ϕ above: $\sigma(x : y : z) = (x^q : y^q : z^q)$. As we shall see in a moment, the points Q and $\sigma(Q)$ have order l , and σ satisfies the equation $\sigma^2 - t\sigma + q = 0$ in the endomorphism ring of $E(T)$. Therefore $t \bmod l$ is characterized by the equality

$$\sigma^2(Q) + q \cdot Q = t \cdot \sigma(Q).$$

Thus, to determine $t \bmod l$ one can simply calculate the left-hand side of this equality, and compare it with $0 \cdot \sigma(Q), 1 \cdot \sigma(Q), 2 \cdot \sigma(Q), \dots$. Here the calculations in $E(T)$ can be done as in §3.

To establish the properties of Q and σ that we used we consider the set V of points $P \in E(K)$ of order l . For each such $P = (x_P : y_P : 1)$ there is a unique \mathbf{F}_q -linear ring homomorphism $T \rightarrow K$ sending $\overline{X}, \overline{Y}$ to x_P, y_P , respectively. It is straightforward to check that the combined ring homomorphism $T \rightarrow \prod_{P \in V} K$ is *injective*, so that $E(T)$ may be considered as a subgroup of $\prod_{P \in V} E(K)$. Since Q corresponds to $(P)_{P \in V}$, it has order l . Also, σ is the restriction to $E(T)$ of the automorphism of $\prod_{P \in V} E(K)$ that on each coordinate is given by ϕ ; hence the equality $\sigma^2 - t\sigma + q = 0$ is a consequence of the equality $\phi^2 - t\phi + q = 0$. Clearly, σ is injective, so $\sigma(Q)$ has order l . This concludes our sketch of Schoof's algorithm.

The algorithm is completely deterministic, and it can be shown to run in time $O((\log q)^8)$. (This is slightly better than Schoof [30], who has $O((\log q)^9)$.) However, it seems that the algorithm is not suited for practical computations.

We remark that Schoof's algorithm does not calculate the structure of the abelian group $E(\mathbf{F}_q)$. It is known that $E(\mathbf{F}_q) \cong \mathbf{Z}/d_1\mathbf{Z} \times \mathbf{Z}/d_2\mathbf{Z}$ for certain positive integers d_1, d_2 for which d_1 divides d_2 , and that d_1 divides

$$\gcd(\#E(\mathbf{F}_q), q-1).$$

V. Miller has shown that if the prime factorization of the latter gcd is known, one can find d_1 and d_2 by means of a probabilistic algorithm that has expected running time $O((\log q)^{c_2})$ for some $c_2 > 0$. For an account of this algorithm, which depends on the *Weil pairing*, we refer to [22].

The fourth method to calculate $\#E(\mathbf{F}_q)$ applies only to curves E that are obtained in a special way. For the sake of simplicity we restrict the discussion to the case that q is a *prime number*.

The *complex multiplication field* of the elliptic curve E over the prime field \mathbf{F}_q is defined to be the field $L = \mathbf{Q}((t^2 - 4q)^{1/2})$, where $t \in \mathbf{Z}$ is such that $\#E(\mathbf{F}_q) = q + 1 - t$. This is an imaginary quadratic field, and its ring of integers

A contains a zero π of the polynomial $X^2 - tX + q$. We have $\pi + \bar{\pi} = t$, $\pi\bar{\pi} = q$, and $\#E(\mathbf{F}_q) = (\pi - 1)(\bar{\pi} - 1)$. This gives an easy way to calculate $\#E(\mathbf{F}_q)$ provided that L is known, which is the case for certain special curves. We illustrate this by means of two examples that were basically known to Gauss. For proofs, see [15, Chapter 18] and also [12, §7; 5].

Let it first be assumed that $q \equiv 1 \pmod{3}$ and that the curve $E = E_{a,b}$ has $a = 0$. Then one can prove that $L = \mathbf{Q}(\sqrt{-3})$. The ring of integers A of L is given by $A = \mathbf{Z}[(1 + \sqrt{-3})/2]$. To find the element $\pi \in A$ with $\#E(\mathbf{F}_q) = (\pi - 1)(\bar{\pi} - 1)$ and $\pi\bar{\pi} = q$ one starts by finding an *ideal* \mathfrak{q} with $A\mathfrak{q} = q\bar{\mathfrak{q}}$, as follows.

One first determines an integer d with $d^2 \equiv -3 \pmod{q}$. This can be done in one of three ways. The first is to apply general zero-finding routines for polynomials over finite fields, see [17, §4.6.2]. The second is to apply a square root extraction algorithm as in [34]. The third is to draw elements $u \in \mathbf{F}_q^*$ until one finds one for which $u^{(q-1)/3} \neq 1$ and to put $d \equiv 2u^{(q-1)/3} + 1 \pmod{q}$. Each of these three methods is probabilistic and practical.

Suppose now that d has been determined. Adding q to d , if necessary, we may assume that d is *odd*. Then $\mathfrak{q} = \mathbf{Z}q + \mathbf{Z}(d + \sqrt{-3})/2$ is a prime ideal of A dividing q , and $q\bar{\mathfrak{q}} = Aq$.

Next one determines an element $\pi \in \mathfrak{q}$ for which $\mathfrak{q} = A\pi$. This can be done by searching for the shortest nonzero vector of \mathfrak{q} , for which there exist standard reduction algorithms. Alternatively, one can calculate $\gcd(q, (d + \sqrt{-3})/2)$ by means of the Euclidean algorithm, which is valid in A . Notice that π is only uniquely determined by \mathfrak{q} up to units of A , of which there are six.

Now let ζ be the unique sixth root of unity in A for which $b^{(q-1)/6} \equiv \zeta \pmod{\mathfrak{q}}$; here b is such that $E = E_{0,b}$. Multiplying π by a suitable sixth root of unity we can achieve that $\pi \equiv \bar{\zeta} \pmod{2\sqrt{-3}}$. Then one has

$$\#E(\mathbf{F}_q) = (\pi - 1)(\bar{\pi} - 1) = q + 1 - 2\operatorname{Re}(\pi).$$

It can be proved that $E(\mathbf{F}_q)$ is isomorphic to $A/(\pi - 1)A$ as an abelian group, so that this method gives the group structure as well.

In the second example that we give we assume that the prime q satisfies $q \equiv 1 \pmod{4}$ and that the curve $E = E_{a,b}$ has $b = 0$. Then one can prove that $L = \mathbf{Q}(i)$ with $i^2 = -1$. It has ring of integers $A = \mathbf{Z}[i]$. As before, one can find a prime ideal \mathfrak{q} of A such that $q\bar{\mathfrak{q}} = Aq$ and an element $\pi \in \mathfrak{q}$ such that $\mathfrak{q} = A\pi$. Denote by ζ the unique fourth root of unity in A for which $(-a)^{(q-1)/4} \equiv \zeta \pmod{\mathfrak{q}}$. Multiplying π by a suitable fourth root of unity we may assume that $\pi \equiv \bar{\zeta} \pmod{2(1+i)}$, and then one has $\#E(\mathbf{F}_q) = (\pi - 1)(\bar{\pi} - 1)$.

We briefly sketch how these results can be generalized to any imaginary quadratic field L . Let A be the ring of integers of L , and denote by j_L the j -invariant of the elliptic curve \mathbf{C}/A over \mathbf{C} (cf. [35, Chapter VI]). It is known that j_L is a zero of an irreducible polynomial $F_L \in \mathbf{Z}[X]$ with leading coefficient 1 and degree equal to the class number of L . Methods to calculate F_L can be found in [37]; see also the last section of [30]. The cases $j = 0$ and $j = 1728$

correspond to the fields $L = \mathbf{Q}(\sqrt{-3})$ and $\mathbf{Q}(i)$ that we just considered; let these now be excluded.

Let q be a prime number that does not divide the discriminant of L , and suppose that $q > 3$. Then there are methods, analogous to those discussed above, to decide whether there exists $\pi \in A$ with $\pi\bar{\pi} = q$, and to find such an element π if it does exist; it is unique up to conjugation and sign. Suppose that indeed π exists. Then it can be shown that the polynomial $(F_L \bmod q) \in \mathbf{F}_q[X]$ splits into distinct linear factors. Denote by j any zero of this polynomial in \mathbf{F}_q . One can prove that $j \neq 0, 1728$. Writing $k = j/(1728 - j) \in \mathbf{F}_q^*$ we now consider the two elliptic curves

$$E = E_{3k, 2k}, \quad E' = E_{3kc^2, 2kc^3}$$

over \mathbf{F}_q , where $c \in \mathbf{F}_q$ is any nonsquare. Then L is the complex multiplication field of each of the two curves E, E' , and the two numbers $\#E(\mathbf{F}_q), \#E'(\mathbf{F}_q)$ are the same as the two numbers $(\pi - 1)(\bar{\pi} - 1), (-\pi - 1)(-\bar{\pi} - 1)$. Presumably there is an easy rule to tell which curve belongs to which number, but I do not know what it is. In practice one can decide between the two cases by picking a point $P \in E(\mathbf{F}_q)$ at random and using that P is annihilated by $\#E(\mathbf{F}_q)$.

This concludes our discussion of the methods to calculate the number of points on an elliptic curve over a finite field.

It is a natural question to ask how the numbers $\#E(\mathbf{F}_q)$ are distributed if q is held fixed and E ranges over all elliptic curves over \mathbf{F}_q , up to isomorphism. In particular, one may ask how often a given number occurs as $\#E(\mathbf{F}_q)$. The answer to the latter question, in terms of class numbers of imaginary quadratic orders, is basically due to Deuring [13]; see also [36, 31]. If q is a *prime* number, then Deuring's result implies that *every* integer of the form $q + 1 - t$ with $|t| < 2\sqrt{q}$ occurs as $\#E(\mathbf{F}_q)$ for some elliptic curve E over \mathbf{F}_q . Moreover, it can be deduced that if E is uniformly distributed over all elliptic curves over \mathbf{F}_q , then $\#E(\mathbf{F}_q)$ is approximately uniformly distributed over the numbers near $q + 1$. More accurately, one has the following proposition, which is useful for the analysis of some of the algorithms to be presented in §§5 and 6.

PROPOSITION. *There are positive effectively computable constants c_3 and c_4 such that for any prime number $q > 3$ and any set S of integers s for which $|s - (q + 1)| < \sqrt{q}$ one has*

$$\frac{\#S - 2}{2[\sqrt{q}] + 1} \cdot c_3(\log q)^{-1} \leq \frac{N}{q^2} \leq \frac{\#S}{2[\sqrt{q}] + 1} \cdot c_4(\log q) \cdot (\log \log q)^2,$$

where N denotes the number of pairs $(a, b) \in \mathbf{F}_q^2$ that define an elliptic curve $E = E_{a,b}$ over \mathbf{F}_q with $\#E(\mathbf{F}_q) \in S$.

Note that N/q^2 is the probability that a random pair (a, b) has the stated property. The proposition asserts that, apart from a logarithmic factor, this probability is essentially equal to the probability that a random number near q is in S .

For the proof of the proposition we refer to [20, Proposition (1.16)].

5. Primality testing. It was first pointed out in [5] and [8] that elliptic curves can be used for primality testing. Goldwasser and Kilian [14] proved, modulo a reasonable assumption, that this leads to a probabilistic primality testing algorithm of which the expected running time is bounded by a constant power of $\log n$, where n is the number to be tested. The algorithm of Goldwasser and Kilian depends on Schoof's method to count the number of points on an elliptic curve (see §4), and for this reason it is currently not of practical value. Atkin [2] developed a variant of this algorithm, in which he employs only the special elliptic curves to which the fourth counting method of §4 applies. His algorithm performs very well in practice, and for the numbers to which it has been applied it beats the method of Adleman et al. [1] as implemented by Cohen and A. K. Lenstra [10]; these numbers have approximately 200 digits. It seems very hard to give an exact running time estimate of Atkin's algorithm; but a rough heuristic analysis indicates that its expected running time is again bounded by a constant power of $\log n$.

All these methods depend on a result similar to the following theorem, which is the analogue of Theorem 1.

THEOREM 2. *Let n be an integer, $n > 1$, with $\gcd(n, 6) = 1$. Let E be an elliptic curve over $\mathbf{Z}/n\mathbf{Z}$, and m, s positive integers with s dividing m . Suppose that there is a point $P \in E(\mathbf{Z}/n\mathbf{Z})$ satisfying*

$$m \cdot P = O,$$

$$\gcd(z_q, n) = 1 \quad \text{for each prime divisor } q \text{ of } s,$$

$$\text{where } m(m/q) \cdot P = (x_q : y_q : z_q).$$

Then $\#E(\mathbf{Z}/p\mathbf{Z}) \equiv 0 \pmod s$ for every prime divisor p of n , and if $s > (n^{1/4} + 1)^2$ then n is prime.

The proof, which is analogous to the proof of Theorem 1, is as follows. Let p be a prime divisor of n , and write $Q = (m/s) \cdot P \in E(\mathbf{Z}/n\mathbf{Z})$. Denote by Q_p the image of Q in $E(\mathbf{Z}/p\mathbf{Z})$. From $m \cdot P = O$ it follows that $s \cdot Q = O$, so the order of Q_p divides s . Also, if q is a prime divisor of s , then $s/q \cdot Q_p = (x_q \bmod p : y_q \bmod p : z_q \bmod p)$. This is not the zero point of $E(\mathbf{Z}/p\mathbf{Z})$, since by hypothesis z_q is not divisible by p . Therefore the order of Q_p is not a divisor of s/q , for any prime q dividing s , so this order is equal to s itself. By Lagrange's theorem it follows that $\#E(\mathbf{Z}/p\mathbf{Z})$ is divisible by s . This proves the first assertion of the theorem. If also $s > (n^{1/4} + 1)^2$ then Hasse's inequality $(p^{1/2} + 1)^2 \geq \#E(\mathbf{Z}/p\mathbf{Z})$ implies that $p > n^{1/2}$, and this can only be true for all primes p dividing n if n is prime. This proves Theorem 2.

The algorithms of Goldwasser-Kilian and Atkin need the above theorem only in the case that s is *prime*, so that only $q = s$ has to be considered in the second hypothesis on P in the above theorem. The following schematic description fits both algorithms.

Let n be a large positive integer that one suspects to be a prime number (cf. the remarks in the introduction). To prove that n is prime one proceeds as follows.

(a) One selects an elliptic curve E over $\mathbf{Z}/n\mathbf{Z}$ and a positive integer m such that the following conditions are satisfied:

- (i) $m < (\sqrt{n} + 1)^2$, and if n is prime then $\#E(\mathbf{Z}/n\mathbf{Z}) = m$;
- (ii) there are integers $k > 1$ and $q > (n^{1/4} + 1)^2$ such that $m = kq$ and such that q is probably prime.

Here *probably prime* means that q passes a pseudoprime test as in [17, p. 379], cf. the introduction. To find *one* pair E, m satisfying (i) and (ii), both the algorithm of Goldwasser-Kilian and Atkin's algorithm generate *many* pairs E, m satisfying (i); we shall see below how this is done. It is then hoped that at least one of these pairs satisfies (ii) as well. To check whether a given pair E, m satisfies (ii), one first subjects m to a factoring algorithm that is efficient in finding small factors, such as trial division, or the Pollard $(p-1)$ -method (see §2), or the elliptic curve method (see §6); next one lets k be equal to the product of the small prime factors of m that are found, and one puts $q = m/k$; finally, one checks whether $k > 1$ and whether q is probably prime in the sense explained above. (Goldwasser-Kilian require that in fact $k = 2$ in (ii); this makes it even easier to check (ii).)

(b) Now suppose that E, m, k, q as in (a) have been found. Then one picks a random point P of the form $(x_P : y_P : 1)$ in $E(\mathbf{Z}/n\mathbf{Z})$. This is done as in the second counting algorithm explained in §4. (This algorithm works if $\mathbf{Z}/n\mathbf{Z}$ is a field, which one believes to be the case; for the algorithm to work it is not necessary that one has a *proof* that $\mathbf{Z}/n\mathbf{Z}$ is a field!) Next one calculates $Q = k \cdot P$. One now hopes that $Q \neq O$; it can be proved that this is the case for more than half of all choices of P , if n is actually prime. If $Q = O$ one picks another point $P \in E(\mathbf{Z}/n\mathbf{Z})$, and one keeps trying until $Q = k \cdot P \neq O$. Suppose now that $Q \neq O$. Then one checks that $q \cdot Q = O$, as must be the case if n is prime (by $q \cdot Q = m \cdot P$ and (i) above). Finally one checks that $\gcd(z, n) = 1$, if $Q = (x : y : z)$; this must also be the case if n is prime, since $Q \neq O$.

(c) The final stage of the algorithm consists of proving that q is prime. This can be done by a recursive application of the algorithm, or, if q is below a certain bound, by a more direct method. Notice that $q = m/k < (\sqrt{n} + 1)^2/2$, so that the depth of the recursion is $O(\log n)$.

If (a), (b), and (c) have been performed successfully, then n is indeed a prime number. This follows from Theorem 2, with $s = q$.

It remains to explain how to find many pairs E, m as in (i). In the Goldwasser-Kilian algorithm this is done as follows. First one draws $a, b \in \mathbf{Z}/n\mathbf{Z}$ at random until $4a^3 + 27b^2 \neq 0$; this happens with probability $(n-1)/n$, if n is indeed prime. Next one checks that $\gcd(n, 4a^3 + 27b^2) = 1$, as should be the case if n is prime. Now one puts $E = E_{a,b}$, and by means of Schoof's algorithm one calculates a number m such that (i) holds. If Schoof's algorithm doesn't work then n is not prime. (If n is not prime, then it is unlikely but not impossible that

Schoof's algorithm calculates a number m ; it is an interesting question which information about n this would provide, and what the significance of m would be.)

Atkin's method to find pairs E, m as in (i) is different. Consider the sequence

$$-3, -4, -7, -8, -11, -15, -19, -20, \dots$$

of discriminants of imaginary quadratic fields; an integer belongs to this sequence if and only if it is negative, not divisible by the square of an odd prime number, and in one of the residue classes $1 \bmod 4, 8 \bmod 16, 12 \bmod 16$. For each Δ in a suitable beginning segment of this sequence, one decides whether the ring of integers $A = \mathbf{Z}[(\Delta + \sqrt{\Delta})/2]$ of the imaginary quadratic field $L = \mathbf{Q}(\sqrt{\Delta})$ contains an element π with $n = \pi\bar{\pi}$, and one finds such an element π if it exists; the probabilistic methods to do this that we referred to in §4 are successful provided that n is prime, but, as above, do not require a *proof* that n is prime. The discriminants for which π does not exist are discarded, and the remaining discriminants Δ each give rise to six (if $\Delta = -3$) or four (if $\Delta = -4$) or two (if $\Delta \leq -7$) pairs E, m as in (i), as explained in §4.

For most values of Δ it is easier to determine the values of m than to calculate the coefficients a, b defining E ; hence, it is wise to test whether m satisfies (ii) before calculating a, b .

This finishes the description of the primality tests of Goldwasser-Kilian and Atkin.

The running time of a suitable version of the Goldwasser-Kilian algorithm can be analyzed with the help of the proposition stated in §4. The result is expressed in the following two theorems. The first one states that if a certain standard conjecture concerning the distribution of primes is true, then the algorithm runs in expected polynomial time. The second theorem asserts that in any case this is true for almost all input primes n .

THEOREM 3. *Suppose that there are positive constants c_5 and c_6 such that for all real numbers $x \geq 2$ the number of primes p with $x \leq p \leq x + \sqrt{2x}$ is at least $c_5\sqrt{x}(\log x)^{-c_6}$. Then on any prime input n , the Goldwasser-Kilian algorithm proves the primality of n in expected time $O((\log n)^{10+c_6})$.*

For the proof we refer to [14]. (The exponent $10 + c_6$ is 1 less than the exponent in [14]. This is due to the corresponding improvement in Schoof's algorithm.)

THEOREM 4. *There exist positive constants c_7 and c_8 such that for all integers $k \geq 2$ the fraction of the set of primes n that have k binary digits and for which the expected running time of the Goldwasser-Kilian algorithm is $\leq c_7(\log n)^{11}$ is at least*

$$1 - c_8 2^{-k^{1/\log \log k}}.$$

For the proof we again refer to [14]. It employs a theorem of Heath-Brown, which states that the hypothesis made in Theorem 3 is true in a certain average sense.

6. Factorization. We describe a method to factor integers that depends on the use of elliptic curves. It is the analogue of Pollard's $(p-1)$ -method described in §2.

Let n be the composite integer that one wishes to factor, and assume that $n > 1$, $\gcd(n, 6) = 1$. Pick a random pair (E, P) , where E is an elliptic curve over $\mathbf{Z}/n\mathbf{Z}$ and $P \in E(\mathbf{Z}/n\mathbf{Z})$. This can be done by choosing $a, x, y \in \mathbf{Z}/n\mathbf{Z}$ at random, putting $P = (x : y : 1)$, and letting E be defined by the pair (a, b) , where b is chosen such that $P \in E(\mathbf{Z}/n\mathbf{Z})$; so $b = y^2 - x^3 - ax$. To be certain that E is an elliptic curve one should check that $\gcd(4a^3 + 27b^2, n) = 1$. As in Pollard's $(p-1)$ -method, one now selects a positive integer k that is divisible by many small prime powers, for example, $k = \text{lcm}\{1, 2, \dots, w\}$ for a suitable bound w . Next one calculates the point $k \cdot P \in E(\mathbf{Z}/n\mathbf{Z})$. This can be done by $O(\log k)$ duplications and additions in the group $E(\mathbf{Z}/n\mathbf{Z})$. If $k \cdot P = (x : y : z)$, one calculates $\gcd(z, n)$. One stops if this gcd is a nontrivial divisor of n . If, on the other hand, this gcd equals 1 or n , then one changes the pair (E, P) and starts all over again. The latter option is not available in Pollard's method.

As for the Pollard $(p-1)$ -method, one can show that a given pair (E, P) is likely to be successful in this algorithm if n has a prime divisor p for which $\#E(\mathbf{Z}/p\mathbf{Z})$ is built up from small primes only. The probability for this to happen increases with the number of pairs (E, P) that one tries.

We refer to [20] for the running time analysis of a variant of the elliptic curve factoring algorithm. Using the proposition from §4 and properties of modular curves one finds an upper bound for the expected running time of the algorithm. This upper bound is expressed in terms of the probability that a random number in the interval $(p+1-\sqrt{p}, p+1+\sqrt{p})$ has all its prime factors below a certain bound, where p denotes the least prime dividing n . To estimate the latter probability we need the following unproved conjecture from analytic number theory.

For a real number $x > e$, define

$$L(x) = e^{\sqrt{\log x \log \log x}}.$$

A theorem of Canfield, Erdős, and Pomerance [7, Corollary to Theorem 3.1] implies the following. Let α be a positive real number. Then the probability that a random positive integer $m \leq x$ has all its prime factors $\leq L(x)^\alpha$ is $L(x)^{-1/(2\alpha)+o(1)}$, for $x \rightarrow \infty$. The conjecture that we need is that the same result is valid if m is a random integer in the interval $(x-\sqrt{x}, x+\sqrt{x})$.

Assuming this conjecture, one arrives at the following running time estimate for the elliptic curve factoring algorithm. Let $n \in \mathbf{Z}$, $n > 1$, be the integer that one wishes to factor, and assume that n is not divisible by 2 or 3 and that it is not a prime power. Let further g be any positive integer. Then the variant of

the elliptic curve factoring algorithm described in [20] finds with probability at least $1 - e^{-g}$ a nontrivial divisor of n within time $gK(p)(\log n)^2$, where p denotes the smallest prime divisor of n and $K: \mathbf{R}_{>0} \rightarrow \mathbf{R}_{>0}$ is a function with

$$K(x) = e^{\sqrt{(2+o(1)) \log x \log \log x}} \quad \text{for } x \rightarrow \infty.$$

The algorithm may be repeated on the divisors that are found, until the complete prime factorization of n is obtained. The conjectural running time estimate will then also contain terms $gK(p')(\log n)^2$ corresponding to the other prime divisors p' of n , with the exception of the largest one. In all cases one may expect the total factoring time to be at most $L(n)^{1+o(1)}$ for $n \rightarrow \infty$, with L as above. The worst case occurs if the second largest prime divisor of n is not much smaller than \sqrt{n} , so that n is the product of some small primes and two large primes that are of the same order of magnitude.

Several other factoring methods have been proposed for which, conjecturally, the running time is $L(n)^{1+o(1)}$ for $n \rightarrow \infty$, such as the class group method [29] and the quadratic sieve [26]; see also the discussion in [11]. However, for these other methods the running time is basically independent of the size of the prime factors of n , whereas the elliptic curve method is substantially faster if the smallest prime factor of n is much smaller than \sqrt{n} .

The storage requirement of the elliptic curve factoring method is only $O(\log n)$. This is also true for the class group method [29], but all other known factoring algorithms of conjectured speed $L(n)^{1+o(1)}$ have a storage requirement that is a positive power of $L(n)$.

We refer to [23, 6] for modifications of the elliptic curve method that improve its practical performance. It turns out that, with these modifications, the elliptic curve method is one of the fastest integer factorization methods that is currently used in practice. The quadratic sieve algorithm still seems to perform better on integers that are built up from two prime numbers of the same order of magnitude; such integers are of interest in cryptography [28].

ACKNOWLEDGMENT. Part of the work on this lecture was done at the University of Chicago. I thank this institution for its hospitality and support.

REFERENCES

1. L. M. Adleman, C. Pomerance, and R. S. Rumely, *On distinguishing prime numbers from composite numbers*, Ann. of Math. (2) **117** (1983), 173–206.
2. A. O. L. Atkin, in preparation.
3. E. Bach and J. Shallit, *Factoring with cyclotomic polynomials* (26th Annual Sympos. Foundations of Comput. Sci. (FOCS), Portland, 1985) IEEE Comput. Soc. Press, Washington, 1985, pp. 443–450.
4. H. Bass, *Algebraic K-theory*, Benjamin, New York, 1968.
5. W. Bosma, *Primality testing using elliptic curves*, Report 85-12, Math. Inst. Universiteit van Amsterdam, 1985.
6. R. P. Brent, *Some integer factorization algorithms using elliptic curves*, Research report CMA-R32-85, The Australian National Univ., Canberra, 1985.
7. E. R. Canfield, P. Erdős, and C. Pomerance, *On a problem of Oppenheim concerning "Factorisatio Numerorum"*, J. Number Theory **17** (1983), 1–28.

8. D. V. Chudnovsky and G. V. Chudnovsky, *Sequences of numbers generated by addition in formal groups and new primality and factorization tests*, Research report RC 11262 (#50739), IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1985.
9. H. Cohen and H. W. Lenstra, Jr., *Primality testing and Jacobi sums*, Math. Comp. **42** (1984), 297–330.
10. H. Cohen and A. K. Lenstra, *Implementation of a new primality test*, Math. Comp. **48** (1987), 103–121 and S1–S4.
11. D. Coppersmith, A. M. Odlyzko, and R. Schroepfel, *Discrete logarithms in $GF(p)$* , Algorithmica **1** (1986), 1–15.
12. H. Davenport and H. Hasse, *Die Nullstellen der Kongruenzzetafunktionen in gewissen zyklischen Fällen*, J. Reine Angew. Math. **172** (1934), 151–182.
13. M. Deuring, *Die Typen der Multiplikatorenringe elliptischer Funktionenkörper*, Abh. Math. Sem. Hansischen Univ. **14** (1941), 197–272.
14. S. Goldwasser and J. Kilian, *Almost all primes can be quickly certified* (Proc. 18th Annual ACM Sympos. on Theory of Computing (STOC, Berkeley, 1986), The Association for Computing Machinery, New York, 1986, pp. 316–329).
15. K. Ireland and M. Rosen, *A classical introduction to modern number theory*, Graduate Texts in Math., vol. 84, Springer-Verlag, New York, 1982.
16. N. M. Katz and B. Mazur, *Arithmetic moduli of elliptic curves*, Princeton Univ. Press, Princeton, N.J., 1985.
17. D. E. Knuth, *The art of computer programming*, vol. 2: *Seminumerical algorithms*, 2nd ed., Addison-Wesley, Reading, Mass., 1981.
18. S. Lang and H. Trotter, *Frobenius distributions in GL_2 -extensions*, Lecture Notes in Math., vol. 504, Springer-Verlag, Berlin, 1976.
19. H. Lange and W. Ruppert, *Complete systems of addition laws on abelian varieties*, Invent. Math. **79** (1985), 603–610.
20. H. W. Lenstra, Jr., *Factoring integers with elliptic curves*, Ann. of Math. (to appear).
21. H. W. Lenstra, Jr. and R. Tijdeman (Editors), *Computational methods in number theory*, Math. Centre Tracts, no. 154/155, Mathematisch Centrum, Amsterdam, 1982.
22. V. S. Miller, *Short programs for functions on curves*, IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1986.
23. P. L. Montgomery, *Speeding the Pollard and elliptic curve methods of factorization*, Math. Comp. **48** (1987), 243–264.
24. H. C. Pocklington, *The determination of the prime and composite nature of large numbers by Fermat's theorem*, Proc. Cambridge Philos. Soc. **18** (1914–16), 29–30.
25. J. M. Pollard, *Theorems on factorization and primality testing*, Proc. Cambridge Philos. Soc. **76** (1974), 521–528.
26. C. Pomerance, *Analysis and comparison of some integer factoring algorithms*, in [21], pp. 89–139.
27. H. Riesel, *Prime numbers and computer methods for factorization*, Progr. Math., vol. 57, Birkhäuser, Boston, 1985.
28. R. L. Rivest, A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Comm. ACM **21** (1978), 120–126.
29. C. P. Schnorr and H. W. Lenstra, Jr., *A Monte Carlo factoring algorithm with linear storage*, Math. Comp. **43** (1984), 289–311.
30. R. J. Schoof, *Elliptic curves over finite fields and the computation of square roots mod p* , Math. Comp. **44** (1985), 483–494.
31. R. J. Schoof, *Nonsingular plane cubic curves over finite fields*, J. Combin. Theory Ser. B (to appear).
32. J. L. Selfridge and M. C. Wunderlich, *An efficient algorithm for testing large numbers for primality*, Proc. Fourth Manitoba Conf. Numerical Math., University of Manitoba, Congressus Numerantium XII, Utilitas Math., Winnipeg, 1975, pp. 109–120.
33. D. Shanks, *Class number, a theory of factorization, and genera*, Proc. Sympos. Pure Math. vol. 20, Amer. Math. Soc., Providence, R.I., 1971, pp. 415–440.

- 34. —, *Five number-theoretic algorithms*, Proc. Second Manitoba Conf. Numerical Math., University of Manitoba, Congressus Numerantium VII, Utilitas Math., Winnipeg, 1973, pp. 51–70.
- 35. J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Math., vol. 106, Springer-Verlag, New York, 1986.
- 36. W. C. Waterhouse, *Abelian varieties over finite fields*, Ann. Sci. École Norm. Sup. (4) **2** (1969), 521–560.
- 37. H. Weber, *Lehrbuch der Algebra*, vol. III, Friedrich Vieweg und Sohn, Braunschweig, 1908; reprinted by Chelsea Publishing Company, New York.
- 38. H. C. Williams, *Primality testing on a computer*, Ars Combin. **5** (1978), 127–185.
- 39. —, *A $p + 1$ method of factoring*, Math. Comp. **39** (1982), 225–234.

MATHEMATISCH INSTITUUT, UNIVERSITEIT VAN AMSTERDAM, 1018 WB AMSTERDAM, THE NETHERLANDS

Recent Progress in Geometric Partial Differential Equations

RICHARD SCHOEN

The equations of differential geometry lead to many natural problems in partial differential equations. Examples of these are the Einstein equations for a metric tensor, the minimal surface equation, the Yang-Mills equations, and the harmonic map equation. In recent years there has been substantial progress made on the analytic theory of basic geometric problems through the use of hard analysis. While the analytic theory has been useful in solving geometric problems, the geometric equations often serve as prototypes which lead to the development of new analytic machinery. For example, the Monge-Ampere equation is a prototype for the fully nonlinear theory which has been developed. In this lecture we will discuss two geometric problems. The first is the Yamabe equation, a conformally invariant scalar equation which arises from conformal deformation of Riemannian metrics. The second is the harmonic map system which arises when one extremizes the energy for maps constrained to lie on a given manifold.

A common feature which is shared by many nonlinear problems in science is the presence of singular behavior. For geometric problems, singularities are often an important aspect of the problem and they contain useful information. In the case of the Yamabe equation, singularities of solutions may be thought of as the points at infinity for a complete, constant scalar curvature, Riemannian metric on a subdomain of the n -sphere. In the harmonic map problem the singular set is the set along which the map tears or is discontinuous. It is an important and challenging problem to understand singularities in solutions of nonlinear equations. We discuss this in more detail in §3.

The resolution of the Yamabe problem (described in more detail below) is the first step in a general program for constructing canonical Riemannian metrics on smooth manifolds. In fact, Yamabe's original goal was to solve the three-dimensional Poincaré conjecture. If M is a smooth compact manifold, then an Einstein metric g on M is a metric whose Ricci curvature is proportional to g , that is $\text{Ric}(g) = cg$ for a constant c . If M is three-dimensional, then an

Einstein metric necessarily has constant curvature. In particular, the existence of an Einstein metric on a simply connected three manifold implies the Poincaré conjecture. A powerful approach to the solution of nonlinear equations is the variational method. The Einstein equation is, in fact, the Euler-Lagrange equation for a natural variational integral, the Einstein-Hilbert action. If we let \mathfrak{m}_1 denote the space of Riemannian metrics g on M satisfying the volume constraint $\int_M dv_g = 1$, then the Einstein-Hilbert action is the functional $\mathcal{R}: \mathfrak{m}_1 \rightarrow \mathbf{R}$ given by

$$\mathcal{R}(g) = \int_M R_g dv_g$$

where R_g is the scalar curvature function of g . Of course the functional $\mathcal{R}(\cdot)$ is invariant under the action of the diffeomorphism group $\text{Diff}(M)$, and hence is defined on $\mathfrak{m}_1/\text{Diff}(M)$. The functional \mathcal{R} is neither bounded above nor below and hence one cannot hope to find either maxima or minima. Analysis of the linearization about a critical metric g shows that, up to finite-dimensional spaces of deformations, \mathcal{R} is minimized within the conformal class of g (the metrics $\bar{g} = e^v g$, $v \in C^\infty(M)$) and is maximized in directions orthogonal to the conformal class of g . This suggests that one may find an Einstein metric by a two-step procedure. First fix a metric $g_0 \in \mathfrak{m}_1$, and consider the minimization problem

$$I(M; g_0) = \inf_{g \in [g_0]} \mathcal{R}(g)$$

where $[g_0]$ denotes the conformal class of g_0 . Now maximize over g_0

$$I(M) = \sup_{g_0 \in \mathfrak{m}_1} I(M; g_0).$$

The first step is the Yamabe problem—that is, the existence of a smooth $g \in [g_0]$ with $\mathcal{R}(g) = I(M; g_0)$. It is perhaps natural that the resolution of the Yamabe problem (described below) relies heavily on ideas from General Relativity which is also the source of the action $\mathcal{R}(\cdot)$. Another consequence of the solution of the Yamabe problem is that $I(M; g_0)$ is uniquely maximized among n -manifolds by the standard S^n . In particular $I(M) \leq I(S^n)$ for any n -manifold M . One may also note that $I(M)$ is positive if and only if M admits a metric with positive scalar curvature. Thus it follows from a result of N. Hitchin [17] that there are exotic spheres M with $I(M) \leq 0$. This shows that $I(\cdot)$ reflects the smooth structure of M . The remaining problem is to realize $I(M)$ by a smooth metric. For a three manifold M , it seems plausible that a maximizing sequence of metrics should converge to a metric which is singular along two-dimensional surfaces. This brings us back to our theme of singular behavior. In this case an understanding of the singularities in this process may be the key to unlocking the mysteries of the manifold.

1. Scalar curvature and conformally invariant scalar equations. We discuss the geometric theory related to the following model equation

$$\Delta u + u^p = 0 \quad (1.1)$$

where u is a positive function defined on an open subset of \mathbf{R}^n for $n \geq 3$. We will assume p is greater than one and less than or equal to $p_0 = (n+2)/(n-2)$; in fact, the case $p = p_0$ is the case of primary interest in geometry. We will discuss the existence theory for classical solutions on a compact manifold as well as the weak solutions $u \in L^p_{\text{loc}}$ defined on \mathbf{R}^n (or S^n). Equations such as (1.1) also arise in many physical contexts. On a smooth bounded domain Ω , the boundary condition $u \equiv 0$ on $\partial\Omega$ has been studied by several authors. Pohozaev [22] showed that if Ω is star-shaped and $p = (n+2)/(n-2)$ then no positive solution exists for this problem. H. Brezis and L. Nirenberg [8] showed that in many cases a positive solution does exist to the modified equation $\Delta u + \lambda u + u^{(n+2)/(n-2)} = 0$ with $0 < \lambda < \lambda_1(\Omega)$. Recently A. Bahri and J. M. Coron [5] have shown that solutions do exist on domains Ω which are not contractible. Their proof involves delicate variational arguments. From a geometric point of view we think of u as defining the Riemannian metric $\bar{g}_{ij} = u^{4/(n-2)}\delta_{ij}$. The equation (1.1) then says that the metric \bar{g} has scalar curvature identically equal to $4(n-1)/(n-2)$. From this interpretation one easily sees the conformal invariance of the equation; since $\gamma^*\bar{g}$ will also have constant scalar curvature we see that if u satisfies (1.1) ($p = p_0$), then $u_\gamma = |\gamma'|^{(n-2)/2}u \circ \gamma$ is also a solution where $|\gamma'|$ denotes the linear stretch factor of the conformal map γ . In the early 1960s H. Yamabe [35] studied a basic existence question for this equation. Yamabe's problem is to construct a Riemannian metric of constant scalar curvature which is conformally related to any given metric g on a compact manifold M^n of dimension n greater than or equal to three. Thus the problem is to produce a positive solution u on M to the equation

$$Lu + Ku^{(n+2)/(n-2)} = 0, \quad Lu = \Delta_g u - \frac{n-2}{4(n-1)}R_g u \quad (1.2)$$

where K is a constant which may be normalized to 1, 0, or -1 . The operator L appearing in (1.2) is referred to as the conformal Laplacian and has a natural invariance under conformal change of metric. The differential equation (1.2) arises as the Euler-Lagrange equation of a variational problem, namely the problem of extremizing the Einstein-Hilbert action

$$\mathcal{R}(\bar{g}) = \int_M R_{\bar{g}} dv_{\bar{g}}$$

subject to the volume constraint $\int_M dv_{\bar{g}} = 1$ where \bar{g} is required to be conformally related to a given Riemannian metric g . Yamabe viewed his paper as a first step toward understanding the critical point theory of $\mathcal{R}(\cdot)$ on the space of Riemannian metrics on M . As described above \mathcal{R} behaves differently along the conformal directions in the space of metrics than it does in orthogonal directions.

If we write $\bar{g} = u^{4/(n-2)}g$, then the variational problem takes the classical form of extremizing

$$\mathcal{R}(\bar{g}) = \frac{4(n-1)}{n-2} \int_M \left(|\nabla u|^2 + \frac{n-2}{4(n-1)} R_g u^2 \right) dv_g$$

subject to the constraint

$$\int_M u^{2n/(n-2)} dv_g = 1.$$

This leads to equation (1.2) for a multiple of u . In 1968 Trudinger [32] found that Yamabe's original paper contained an error, and he proved the existence for (1.2) in the case $R_g \leq 0$. (It is no loss of generality to assume that R_g is of one sign.) The next major progress was made by T. Aubin [3] in 1975 when he constructed a minimizing function u in case $n \geq 6$ and g is not locally conformally flat. The hypotheses of Aubin's theorem were used to derive a correction term in the problem arising from the local conformal distortion of g . An earlier result of Pohozaev [22] made it clear that the existence problem could not be purely local if g is a locally conformally flat metric. It seemed that the missing ingredient in the Yamabe problem should come from global considerations.

From another direction, the study of Riemannian manifolds of nonnegative scalar curvature arises in General Relativity. The local positivity of matter density implies, via the Einstein equation, a curvature inequality on the spacetime metric. Restricting to a spacelike hypersurface M we get a scalar inequality on the Riemannian manifold M . In case M is a maximal hypersurface in the spacetime this inequality implies nonnegative scalar curvature. The standard boundary condition which is imposed on the spacetime is that it be asymptotically flat. In fact, if M is maximal, it is reasonable to require the asymptotic condition

$$g_{ij} = \left(1 + \frac{E}{|x|} \right) \delta_{ij} + O(|x|^{-2})$$

for a coordinate chart near infinity on M . The number E is then the ADM energy [2] and is interpreted as the total energy for the gravitating system measured at spatial infinity. A special case of the positive energy theorem then says that $E \geq 0$ and $E = 0$ only if (M, g) is isometric to \mathbf{R}^3 with the Euclidean metric. The positive energy theorem was demonstrated by the speaker and S. T. Yau [26] in the late seventies. Our proof was based on a study of area minimizing surfaces in M . An interesting proof of the positive energy theorem using the Dirac operator was found by E. Witten [34] a few years later. Recently we used an extension of our minimal surface arguments to prove a singularity theorem in general relativity [27]. This was a statement to the effect that a high concentration of matter necessarily implies the existence of a space-time singularity. There is a natural n -dimensional analogue of the theorem mentioned above where one may require the asymptotic condition

$$g_{ij} = \left(1 + \frac{E}{|x|^{n-2}} \right) \delta_{ij} + O(|x|^{1-n}).$$

More generally it has been shown by R. Bartnik [6] and Lee-Parker [18] that the energy is well defined provided $g_{ij} = \delta_{ij} + o(|x|^{-(n-2)/2})$. The n -dimensional positive energy theorem then says that if $R_g \geq 0$ on M and g is asymptotically flat then $E \geq 0$. Moreover $E = 0$ only if (M, g) is isometric to the flat \mathbf{R}^n . This theorem can be proved using methods analogous to those used in the three-dimensional case.

We now return to the case of a compact M with positive scalar curvature, the setting for the Yamabe problem. Since $R_g > 0$, the operator L is negative definite, and hence for any point $x_0 \in M$ there exists a positive fundamental solution G_{x_0} for L with pole at x_0 . If we let $\bar{g} = G_{x_0}^{4/(n-2)} g$ be the metric determined by G_{x_0} , then we see that $(M - \{x_0\}, \bar{g})$ is an asymptotically flat manifold with $R_{\bar{g}} \equiv 0$. In the case that $n < 6$ or, for general n , if g is locally conformally flat near x_0 the asymptotic behavior of \bar{g} is good enough so that the energy $E(x_0)$ is well-defined. The positive energy theorem then says that $E(x_0) \geq 0$ and $E(x_0) = 0$ only if $(M - \{x_0\}, \bar{g})$ is isometric to Euclidean space. This second possibility can occur only if (M, g) is conformally diffeomorphic to the standard S^n . The speaker [24] has shown that the positivity of $E(x_0)$ for some x_0 provides the global correction term which can be used to construct a minimizing function u for the Yamabe problem. This covers all remaining cases and hence completes the solution of the Yamabe problem.

The analysis of the conformal Laplacian L , in particular its fundamental solution, has also given information about a class of locally conformally flat manifolds including those of nonnegative scalar curvature. We briefly describe this joint work with S. T. Yau [28]. The simplest conformally flat manifold is a domain $\Omega \subset S^n$ in the n -sphere. There is also a class of conformally flat manifolds which are quotients of a domain Ω by a discrete subgroup Γ of the conformal group $\mathcal{G}_n = O(n+1, 1)$ of S^n . If we assume that Γ preserves Ω and acts properly discontinuously and without fixed point on Ω , then $M = \Omega/\Gamma$ is a conformally flat manifold. We refer to such discrete groups Γ as Kleinian groups. A general conformally flat structure on a manifold M is given by an immersion $\Phi: \tilde{M} \rightarrow S^n$ where \tilde{M} is the universal covering of M together with a homomorphism $\rho: \pi_1 M \rightarrow \mathcal{G}_n$ satisfying the equivariance property $\Phi \circ \gamma = \rho(\gamma) \circ \Phi$ for $\gamma \in \pi_1 M$. The map Φ is called the developing map and ρ is called the holonomy representation. If the map Φ is injective, then letting $\Omega = \Phi(\tilde{M})$ and $\Gamma = \rho(\pi_1 M)$ we find that Γ is discrete and $M = \Omega/\Gamma$ as above. Assume for simplicity that M is compact and locally conformally flat, and define a number $d(M)$ as follows: Choose a compatible metric g on M , let $x_0 \in \tilde{M}$ and let G_{x_0} be the minimal Green's function for L with pole at x_0 . We then define

$$d(M) = \inf \left\{ \frac{n-2}{2} p: G_{x_0} \in L^p(\tilde{M} \setminus B_1(x_0)) \right\}.$$

(Note that one can show that G_{x_0} always exists.) It can be seen that $d(M)$ lies in the interval $[0, n]$ and that $d(M)$ is independent of the choice of metric g hence depends only on the conformally flat structure. The existence of a compatible

metric g on M with $Rg > 0$ is equivalent to the inequality $d(M) < (n-2)/2$. The main structure theorem is that if $d(M) < (n-2)^2/n$, then Φ is injective and hence $M = \Omega/\Gamma$ for a Kleinian Γ . Moreover, in the Kleinian group case the number $d(M)$ is the Hausdorff dimension of the limit set $\Lambda = S^n - \Omega$ of Γ . In particular, we see that locally conformally flat manifolds M with positive scalar curvature are precisely of the form Ω/Γ where the limit set of the Kleinian group Γ has Hausdorff dimension less than $(n-2)/2$. The proof of the main theorem involves showing the injectivity of the developing map Φ . This is done by a (somewhat nonstandard) Bochner argument. An easy consequence of the structure theorem is a direct proof of the positive energy theorem for locally conformally flat manifolds with dimension $n \geq 4$.

We now return full circle to the model equations (1.1) and (1.2). A function $u \in L^{(n+2)/(n-2)}(S^n)$ is a (global) weak solution of (1.2) if for every $\phi \in C^\infty(S^n)$ we have

$$\int_{S^n} [(L\phi)u + \phi u^{(n+2)/(n-2)}] dv_0 = 0.$$

A similar definition holds for L^p_{loc} weak solutions of (1.1). Weak solutions of these equations have been considered by many authors including J. Serrin, B. Gidas, J. Spruck, P. Aviles, and L. Caffarelli [29, 13]. A complete understanding of the asymptotic behavior of a weak solution near an isolated singularity has been given by these authors. The theory of weak solutions of (1.2) is related to geometry by the following theorem. Suppose $\Omega \subset S^n$ and u is a classical solution on Ω of (1.2) such that the Riemannian metric $u^{4/(n-2)}g_0$ is geodesically complete on Ω . Then the complement of Ω has Hausdorff dimension at most $(n-2)/2$, u lies in $L^{(n+2)/(n-2)}(S^n)$, and u is a global weak solution of (1.2). Of course the singular set of u is precisely the set $S^n - \Omega$. A consequence of this result together with the solution of the Yamabe problem on compact locally conformally flat manifolds is the construction of a large family of new global weak solutions of (1.2). These are given by complete metrics on the universal covering domain Ω of the manifolds. These weak solutions tend to have singular sets which are of fractional Hausdorff dimension; for example Cantor sets or quasi-spheres. We conjecture that any positive global weak solution of (1.2) arises from a complete metric on a subdomain of S^n .

CONJECTURE. *If $u \in L^{(n+2)/(n-2)}(S^n)$ is a weak solution of (1.2), then u is regular on a domain $\Omega \subset S^n$ and the metric $u^{4/(n-2)}g_0$ is geodesically complete on Ω .*

An intriguing question in the theory of weak solutions is the existence problem. From a geometric point of view this is the problem of constructing a complete conformal metric of constant positive scalar curvature on a given domain $\Omega \subset S^n$. This would give a global weak solution of (1.2) which is singular precisely on $S^n - \Omega$. A necessary condition on Ω is that the Hausdorff dimension of $S^n - \Omega$ be less than $(n-2)/2$. For complete metrics of constant negative scalar curvature the existence problem was solved in reasonable generality by Loewner and

L. Nirenberg [19]. This case is different in that the solutions on Ω do not extend as weak solutions of the corresponding equation. A uniqueness theorem for smooth global solutions of (1.2) was proved by M. Obata [21] and B. Gidas, W. M. Ni, and L. Nirenberg [12]. The methods were extended [13] to show that no global weak solution of (1.2) exists with precisely one singular point on S^n . Recently the speaker has constructed global weak solutions of (1.2) which are singular precisely at a specified finite collection of at least two points on S^n . An interesting general question about weak solutions is whether the singular set necessarily has any special local geometric structure.

2. The harmonic map problem. We describe very briefly those aspects of the harmonic map problem related to weak solutions and regularity theory. Let M^n and N^k be Riemannian manifolds and assume N is isometrically embedded in \mathbf{R}^K . Let $\Psi(u) = (\psi_1(u), \dots, \psi_{K-k}(u))$ be a nonsingular vector of functions defining N , so that $N = \{u \in \mathbf{R}^K : \psi(u) = 0\}$. A map u from M to N is then given by an \mathbf{R}^K -valued vector function satisfying $\Psi(u(x)) = 0$ for $x \in M$. We consider the Dirichlet integral

$$E(u) = \sum_{i=1}^K \int_M |\nabla u_i|^2 dv$$

for maps $u: M \rightarrow N$. The critical maps are then the harmonic maps and they satisfy the Euler-Lagrange equation

$$\Delta u^i + A_{u(x)}^i(\nabla u(x), \nabla u(x)) = 0, \quad i = 1, \dots, K \quad (2.1)$$

where A is the vector-valued second fundamental form of N in \mathbf{R}^K . The harmonic map problem has been much studied in geometry. It is also a common problem in the mathematical physics literature. Recently it has been studied as a simplified model for liquid crystals (see [15]). In order to be brief we refer the listener to survey papers [9, 23] for general discussion and we summarize the regularity aspects of the problem. The existing analytic theory has been done by C. B. Morrey [20], Eells-Sampson [10], Hildebrandt-Kaul-Widman [16], Giaquinta-Giusti [11], Schoen-Uhlenbeck [25], and Hardt-Kinderlehrer-Lin [15]. There is no existing regularity theorem for weak solutions of (2.1). For maps u which are stationary points for the variational problem (see [23] for discussion) regularity for $n = 2$ was shown by M. Gruter [14] and the speaker [23]. It is a conjecture that partial regularity holds for stationary u with $n \geq 3$. The minimizing case is much better understood. We briefly describe the results of Schoen-Uhlenbeck [25] which is the only regularity theory for harmonic maps ($n \geq 3$) not requiring special hypotheses on N . The prototype for a singularity arises from a nonconstant harmonic map $u_0: S^k \rightarrow N$ where $k \geq 2$. Defining $u(x) = u_0(x/|x|)$ then gives a homogeneous harmonic map u from \mathbf{R}^{k+1} into N . If u is locally energy minimizing in \mathbf{R}^{k+1} we refer to u as a minimizing tangent map. For a general minimizing map u we let $S(u)$ denote its singular set. The regularity theorem says that $S(u)$ has Hausdorff dimension at most

$n - 3$ in general. If no nontrivial MTM's exist into N for $k = 2, \dots, l$ then $\dim S(u) \leq n - l - 1$. Moreover, the theorem shows that for any $P_0 \in S(u)$ we get a nontrivial MTM (possibly singular away from the origin) by taking limits of dilated maps $u_{\lambda_i}(x) = u(\lambda_i x)$ for a normal coordinate system x centered at p_0 and a sequence λ_i tending to zero. This explains the terminology "tangent map."

3. Remarks on singularities. Differential geometry provides several natural nonlinear partial differential equations for which singular behavior of solutions is an intrinsic part of the problem. Besides the two examples we have mentioned one should mention the Plateau problem and the hyperbolic Einstein equation. In some instances conjectures have been made concerning the nature of the singularities such as the cosmic censorship conjecture of R. Penrose for the Einstein equations.

There is a collection of questions or conjectures existing for the singular structure of the minimal surface problem and the harmonic map equation. I will do the discussion in the harmonic map context although much of it would be analogous for minimal varieties. The questions I will discuss have been posed over the last twenty years primarily in the context of minimal varieties. I have chosen to discuss harmonic maps because the theory of weak solutions is easier to describe. There are three basic areas in which questions have been posed. These are concerning (1) the structure of the singular set, (2) stability of singularities, and (3) prescribing singularities.

The main conjecture in (1) for a harmonic map is that $S(u)$ is a union of integer dimensional rectifiable strata. This conjecture seems very plausible under the hypothesis that the manifolds M, N are real analytic Riemannian manifolds. It is very far from being known and one could easily encounter disagreement among experts as to whether it is likely to be true. One instance where the structure of the singular set has been identified is in the work of J. Taylor [31] on two-dimensional area minimizing sets in three space. In this case the singular set is a one-dimensional complex. A major obstruction in proving anything about $S(u)$ is the question of uniqueness of the tangent map at a singular point $p_0 \in S$. Roughly speaking, nonuniqueness would imply that the behavior of the map is very different at different scales and has no limiting configuration under higher magnification. A recent breakthrough on this problem was made by L. Simon [30] who proved tangent map (and tangent cone for minimal varieties) uniqueness in the isolated singularity case assuming that N is real analytic. The uniqueness under an additional hypothesis had been proven by W. Allard and F. Almgren [1]. For nonisolated singularities the uniqueness is still unknown. The point of (2) is the question of whether the singular structure changes drastically under perturbation of the problem, either boundary data or metric perturbation. There are instances when it would be useful to know that singularities disappear entirely under a small perturbation of the problem. Some progress has been made on this for two-dimensional minimal varieties by R. Böhme and

A. Tromba [7] and B. White [33]. Many open questions remain. For the harmonic map problem it does not seem reasonable to attempt to specify both boundary data and singular set as we did for the scalar equations of part one of our lecture. One may hope to specify a singular set S and ask for the local existence of a weak solution singular on S . The known examples of singularities are relatively few so it would be useful to have a better idea of what can happen locally. Work in this direction has been done by L. Caffarelli, R. Hardt, and L. Simon.

If we analyze these same three areas for weak solutions of the conformally invariant scalar equation of part one, we see that the singular behavior is very different from the expected harmonic map behavior. We have pointed out that $S(u)$ can be a set of fractional Hausdorff dimension even for a global solution. While this is fairly natural from the geometry of the situation, it is perhaps surprising behavior for solutions of such a simple analytic elliptic equation. A partial technical explanation for this difference is the lack of a *monotonicity inequality* for the scalar equation. The harmonic map regularity theory and also the work of L. Simon on tangent map uniqueness rely heavily on the fact that the scale invariant energy for a harmonic map is monotone. Roughly speaking this says that the average behavior of the map tends to improve at higher magnification. On the other hand, solutions of the scalar equation exist which are invariant under a sequence of higher and higher magnifications but which are not homogeneous. This would seem to prohibit the possibility of a monotonicity inequality. In conclusion, there are many interesting singularity phenomena arising from geometric problems which deserve more study, and this should be a fruitful direction for future research.

REFERENCES

1. W. Allard and F. J. Almgren, *On the radial behavior of minimal surfaces and the uniqueness of their tangent cones*, Ann. of Math. **113** (1981), 215–265.
2. R. Arnowitt, S. Deser, and C. Misner, *Energy and the criteria for radiation in general relativity*, Phys. Rev. (2) **118** (1960), 1100.
3. T. Aubin, *The scalar curvature*, Differential Geometry and Relativity (Cahen and Flato, eds.), Reidel, Dordrecht, 1976.
4. P. Aviles, *On isolated singularities in some nonlinear partial differential equations*, Indiana Univ. Math. J. **32** (1983), 773–791.
5. A. Bahri and J. M. Coron, *Une théorie des points critiques à l'infini pour l'équation de Yamabe et le problème de Kazdan-Warner*, C. R. Acad. Sci. Paris Sér. I Math. **15** (1985), 513–516.
6. R. Bartnik, *The mass of an asymptotically flat manifold*, Preprint.
7. R. Böhme and A. Tromba, *The index theorem for classical minimal surfaces*, Ann. of Math. **113** (1981), 447–499.
8. H. Brezis and L. Nirenberg, *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents*, Comm. Pure Appl. Math. **36** (1983), 437–477.
9. J. Eells and L. Lemaire, *A report on harmonic maps*, Bull. London Math. Soc. **10** (1978), 1–68.
10. J. Eells and J. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math. **86** (1964), 109–160.

11. M. Giaquinta and E. Guisti, *On the regularity of the minima of variational integrals*, Acta Math. **148** (1982), 31–46.
12. B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys. **68** (1979), 209–243.
13. B. Gidas and J. Spruck, *Global and local behavior of positive solutions of nonlinear elliptic equations*, Comm. Pure Appl. Math. **34** (1981), 525–598.
14. M. Grüter, *Conformally invariant variational integrals and the removability of isolated singularities*, Manuscripta Math. **47** (1984), 85–104.
15. R. Hardt, D. Kinderlehrer, and F. H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Comm. Math. Phys. (to appear).
16. S. Hildebrandt, H. Kaul, and K. O. Widman, *An existence theorem for harmonic mappings of Riemannian manifolds*, Acta Math. **138** (1977), 1–16.
17. N. Hitchin, *Harmonic spinors*, Adv. in Math. **14** (1974), 1–55.
18. J. Lee and T. Parker, Preprint.
19. C. Löwner and L. Nirenberg, *Partial differential equations invariant under conformal or projective transformations*, Contributions to Analysis: A collection of Papers Dedicated to Lipman Bers, Academic Press, New York, 1974.
20. C. B. Morrey, *The problem of Plateau on a Riemannian manifold*, Ann. of Math. **49** (1948), 807–851.
21. M. Obata, *The conjectures on conformal transformations of Riemannian manifolds*, J. Differential Geom. **6** (1971), 247–258.
22. S. Pohozaev, *Eigenfunctions of the equation $\Delta u + \lambda f(u) = 0$* , Soviet Math. Dokl. **6** (1965), 1408–1411.
23. R. Schoen, *Analytic aspects of the harmonic map problem*, Seminar on Nonlinear Partial Differential Equations, MSRI Publications 2, Springer-Verlag, 1984.
24. —, *Conformal deformation of a Riemannian metric to constant scalar curvature*, J. Differential Geom. **20** (1984), 479–495.
25. R. Schoen and K. Uhlenbeck, *A regularity theory for harmonic maps*, J. Differential Geom. **17** (1982), 307–335.
26. R. Schoen and S. T. Yau, *On the proof of the positive mass conjecture in General Relativity*, Comm. Math. Phys. **65** (1983), 45–76.
27. —, *The existence of a black hole due to condensation of matter*, Comm. Math. Phys. **90** (1983), 575–579.
28. —, *Conformally flat manifolds, scalar curvature, and Kleinian groups*, Preprint.
29. J. Serrin, *Removable singularities of solutions of elliptic equations. II*, Arch. Rational Mech. Anal. **20** (1965), 163–169.
30. L. Simon, *Asymptotics for a class of non-linear evolution equations, with applications to geometric problems*, Ann. of Math. **118** (1983), 525–571.
31. J. Taylor, *The structure of singularities in soap-bubble-like and soap-film-like minimal surfaces*, Ann. of Math. **103** (1976), 489–539.
32. N. Trudinger, *Remarks concerning the conformal deformation of Riemannian structures on compact manifolds*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3) **22** (1968), 265–274.
33. B. White, *Generic regularity of unoriented two-dimensional area minimizing surfaces*, Ann. of Math. **121** (1985), 595–603.
34. E. Witten, *A new proof of the positive energy theorem*, Comm. Math. Phys. **80** (1981), 381–402.
35. H. Yamabe, *On a deformation of Riemannian structures on compact manifolds*, Osaka J. Math. **12** (1960), 21–37.

UNIVERSITY OF CALIFORNIA, SAN DIEGO, LA JOLLA, CALIFORNIA 92093, USA

Equation Solving in Terms of Computational Complexity

ARNOLD SCHÖNHAGE

1.

1.1. *Introduction.* Computational complexity is a new principle in Mathematics, rooted in the algorithmic and constructive tradition of our science. Beyond its prominent role in theoretical Computer Science this aspect has meanwhile entered many different areas of Mathematics. Complexity considerations are intimately related to Logic and Foundations and to Numerical Methods with their innumerable applications, but they are also of growing interest in other fields like Geometry, Number Theory, and Algebra.

In order not to get stuck in pure generality, we confine our very broad subject to the computational treatment of *algebraic* equations. Even in this restricted sense, though, “equation solving” has been investigated for centuries and is now a central topic of Numerical Mathematics and in Computer Algebra. The specific interest of this survey is in the computational *complexity* of such problems. Beginning with a period of a few early papers, corresponding research has been carried on continuously for about twenty years now. Thus it seems to be timely to report on some of the results on this occasion.

In view of limited space and time we will discuss *sequential* algorithms only. A survey of important complexity results for models of *parallel* computation has recently been given by Cook [9]. According to personal taste and preference, we will furthermore restrict our considerations to *deterministic* computations, not ignoring the fact that, from a practical point of view, *probabilistic* algorithms may prove to be superior for certain difficult problems. Moreover, there is always a natural interplay between equation solving and the nondeterministic mode of solution guessing. Accordingly, it will sometimes be illuminating to compare the complexity of equation *solving* with that of *verification*, which means checking whether given data are really forming a solution.

Solving algebraic equations can be understood in different ways. For the prime fields $GF(p)$ and \mathbb{Q} together with their finite extensions the discrete methods of symbolic computation apply. In this framework the major task may be defined as the construction of suitable splitting fields. We mention two papers, [23] and [6],

dealing with the complexity of the general problem. With regard to applications, however, the upper bounds obtained by these authors are discouragingly high. It is possibly premature to expect complete answers while we are still unable to master the complexity bounds for solving systems of *linear* equations (see §4). Considerable progress has been made in factoring polynomials and on related topics. In their pioneering paper [26], A. K. Lenstra, H. W. Lenstra, and L. Lovász have shown how to factor univariate polynomials over \mathbf{Q} in polynomial time. Meanwhile many variations, extensions, refinements of this result, or of the underlying basis reduction algorithm, have been published (see [5, 13, 21, 24, 25, 39] and the extensive bibliography in [15]). In particular, testing for *solvability* in the very classical sense of Galois theory is possible in polynomial time, cf. [22].

When dealing with equations over \mathbf{R} or \mathbf{C} , we should adopt a different point of view. Our complexity results about the fundamental theorem of algebra presented in §3 will refer to computations with dynamic precision in the sense of recursive dependence, quite in the spirit of H. Weyl's farseeing paper [53] of 1924. More specific explanations about the corresponding model of computation are given below in §1.2.

Within the last years, the complexity of root finding and related problems have independently been investigated by S. Smale and others, but along completely different lines so that their results are incomparable to ours. In view of the program of this Congress it should perhaps suffice just to mention [43] and the overview [44] with its further references.

The subsequent presentation deals with rather basic problems of equation solving. §2 on the complexity of multiplication and division of integers and polynomials can be regarded as a complexity discussion of the primitive equation $ax = b$ over corresponding domains. In §3 we present the fundamental theorem of algebra in terms of computational complexity, including a neat solution for this long standing numerical problem. The final section will concern systems of linear equations and the complexity of matrix computations, characteristic equations, and the computation of eigenvalues.

1.2. Models of computation. There are many equivalent ways to define computability. Among these, the classical Turing machine concept is especially suited to provide intuitively appealing measures for the quantitative analysis of algorithms. As long as we are only interested in complexity classes invariant under polynomial reducibilities, further details about the underlying model do not matter. In lower order complexity, however, finer distinctions should be made. Many of the concrete algorithms given in the literature can be made to work for *multitape* Turing machines. For higher flexibility, *random access machines* with a variety of instruction sets can be used, conveniently modeling the addressable storage features of present-day computers. In some sense, the "true," canonical basis for measuring time-complexity is furnished by the theoretical model of *pointer machines* (previously also called "linking automata," or "storage modification machines," cf. [35]). They are real-time equivalent to very simple random

access machines under unit cost, only capable of indirect addressing, testing for equality, and computing the successor function.

With respect to practical implementations, it will be more realistic to consider random access machines under the *logarithmic cost criterion*, where handling any integer is weighted by the length of its binary code. Then any upper bound T on “pointer time” (the number of steps of some pointer machine under consideration) will get amplified by not more than a factor of order $\log T$. One should keep in mind, however, that the applicability of this model is limited by the size of inner storage. For large scale problems with higher storage requirements, the multitape Turing machines may form the more adequate model again.

When discussing time complexity for such machine models, some convention about the encoding of input and output data is required. For simplicity, let us assume here that *integers* shall always be represented in standard *binary* form, though it cannot be excluded that other encodings may lead to different results (as an example, testing for powers of ten seems to be much easier for inputs in decimal form). Elements of $GF(p)$ are given as (reduced) residues mod p , rational numbers as pairs of integers, and elements from finite extensions of these domains are then representable as tuples of binary integers in a straightforward manner.

For approximate computations with real or complex numbers it is especially convenient to use *binary rationals*, i.e., binary integers scaled with some power of two. There is, however, a significant difference between the ad hoc notion of floating point numbers (called “reals” in everyday programming languages) and our understanding of a real number. Any *input* number $\alpha \in \mathbb{R}$ shall (potentially) be available at any desired precision: when called with a specified parameter value N , some *oracle* will deliver some binary rational a such that $|\alpha - a| < 2^{-N}$ is satisfied, without extra cost. Different oracles may possibly specify the same real number in this way. Despite the availability of arbitrary high precision for inputs α, β their equality $\alpha = \beta$ is recursively undecidable. Similarly, outputs resulting from an approximating computation will be due up to some prescribed precision again.

In the abstract setting of *algebraic complexity theory* (see [4] and the more recent survey [50]) one uses the machine independent notion of *straight-line programs*. Usually, the inputs are considered as indeterminates x_1, x_2, \dots over some ground field F , and time is measured (sequentially) by the number of arithmetical operations carried out by such a program in $F(x_1, \dots)$, or in $F[x_1, \dots]$, if divisions are not admitted. *Nonscalar complexity* refers to the mode where only *essential* multiplications and divisions are counted, while additions and scalar multiplications are taken for free. In case of problems for which comparison based branching (or zero testing) is required the more general model of computation trees applies (cf. [49]).

2.

2.1. *The complexity of multiplication and division.* Let us begin with the very simple equation $ax = b$. When considered over the integers, testing its solvability

and finding a solution x is possible by integer division with remainder zero. For checking a given solution x , integer *multiplication* will suffice, but it is an interesting open problem whether there are faster methods of verification. Over the rationals (represented by pairs of integers), solving $ax = b$ amounts to two integer multiplications, followed by a gcd computation provided the result shall be delivered in reduced form. Similarly, an *extended* gcd computation will be required for solving $ax = b \bmod q$ (assuming that b is divisible by $d = \gcd(a, q)$), i.e., to find the gcd d and *cofactors* u, v satisfying the conditions

$$au + qv = d \quad \text{and} \quad d|a \quad \text{and} \quad d|q. \quad (2.1)$$

The key to all these tasks is approximate division over the reals. Its complexity can be bounded by that of integer multiplication, since computing an n -bit approximation for $1/a$ (for $1 \leq a < 2$, say) is possible by means of a division-free Newton iteration with two multiplications per step, whence the time complexity of n -bit division is bounded by

$$O(\mu(n) + \mu(n/2) + \mu(n/4) + \mu(n/8) + \cdots) \leq O(\mu(n)), \quad (2.2)$$

where $\mu(N)$ denotes any bound for the time complexity of N -bit integer multiplication (satisfying some regularity condition to justify the estimation of the sum in (2.2)). Therefore, solving $ax = b$ over the reals (with a bounded away from zero) has essentially no higher complexity than verification of a solution by multiplication. Conversely, $ab = b/(1/a)$ shows how to reduce multiplications to divisions.

In [18, §4.3.3], D. Knuth gives a rather complete account of what is presently known about the time complexity of integer multiplication. For multitape Turing machines, the upper bound from our 1971 paper [41] based on FFT methods modulo numbers of the form $2^L + 1$ is still the best we have. A streamlined version of this approach is contained in [37]. On the other hand, so far nobody has derived any nonlinear *lower* bound, except for models under certain additional restrictions. In fact, any such lower bound proof would necessarily have to employ some specific properties of Turing machines, since pointer machines, for instance, are capable of integer multiplication in *linear time* (cf. [35]). The latter result is based upon the numerical FFT with suitable precision. Accordingly, our present knowledge about the time complexity of solving the equation $ax = b$ over \mathbf{Z} , \mathbf{R} (or over \mathbf{C} as well) is captured by the following bounds.

(2.3) Upper bounds for the time complexity of N -bit multiplication:

$$\mu(N) = cN \log(N+1) \log \log(N+2) \quad \text{for multitape Turing machines,}$$

$$\mu(N) = cN \quad \text{for pointer machines (unit cost),}$$

$$\mu(N) = cN \log(N+1) \quad \text{for pointer machines (logarithmic cost).}$$

For the other cases mentioned above, where gcd computations come in, the complexity seems to be higher. Based on Knuth's idea to combine fast integer multiplication with a "half-gcd" technique due to Lehmer, it has been shown [31] that computing the gcd of two numbers of at most N bits in length (together

with corresponding cofactors) is possible in time $O(\mu(N) \log N)$ (opposed to time of order N^2 for Euclid's algorithm). It seems to be a rather safe conjecture that the extra factor of order $\log N$ over multiplication time is inevitable. Note, however, that, for given cofactors u, v , *verification* of (2.1) is indeed possible within the order of multiplication time.

In the framework of algebraic complexity with unit cost per arithmetical operation, the complexity of solving $ax = b$ seems to be a trivial subject, as one division will suffice, but things become highly nontrivial, if we discuss this problem of equation solving for finite-dimensional algebras over some field F . Consider, for instance, the field of complex numbers as an algebra over \mathbf{R} . One can fairly easily see that the nonscalar complexity of multiplying two complex numbers equals 3 (counting real multiplications and divisions only), but it was only recently that the corresponding problem for the *division* of complex numbers could be settled; cf. [27] for the rather intricate proof that the known upper bound of 6 real multiplications/divisions is indeed optimal.

Much work has been done concerning the multiplicative complexity of algebras. For further details and corresponding references we recommend the survey [50]. Less information is available about "division" in this framework. In the next section we shall briefly discuss the special case of the division of univariate polynomials. The central topic of matrix inversion will be postponed to §4.3.

2.2. Basic computations with polynomials. Now we consider the equation $a(t)x(t) = b(t)$ for polynomials over some field F , at first with regard to the algebraic model counting all arithmetical operations in F at unit cost. Input and output shall be coefficientwise, with dense encoding. More precisely, F is assumed to have the form $F = G(a_0, a_1, \dots, b_0, \dots)$ with all the inputs as indeterminates over some ground field G . Again we may distinguish various domains like $F[t]$, $F(t)$, or $F[t]/(q(t))$, in analogy to \mathbf{Z} , \mathbf{Q} , $\mathbf{Z}/q\mathbf{Z}$ discussed before. Here the role of \mathbf{R} is taken by the ring of formal power series over F . Operating with variable precision in $F[[t]]$ means to operate in $F[t]/(t^m)$ for increasing values of m . Thus the complexity of the division of polynomials can be reduced to that of polynomial multiplication (again up to a constant factor) via computing approximate reciprocals of units in $F[[t]]$ by means of Newton iteration, see [42, 19]. Actually, these authors are discussing the simpler case of *nonscalar* complexity, which, for polynomial multiplication mod t^m , is exactly known ($= 2m - 1$, provided the ground field G contains at least $2m - 2$ elements). Kung's upper bound of $4m$ for computing reciprocals mod t^m and its conjectured optimality can be replaced by the better estimate $3.75m$, while the precise constant factor is yet unknown.

The algebraic complexity of polynomial multiplication is essentially that of discrete convolutions, thus closely related to discrete Fourier transforms. If the ground field G contains suitable roots of unity, e.g. all 2^k th or all 3^k th roots of 1 (for all k), then FFT can be used immediately; otherwise extra measures are required. The best bounds known so far are very similar to those for integer multiplication.

(2.4) Upper bounds for the number of arithmetical operations for the multiplication of m th degree polynomials over ground field G :

$$\begin{aligned} M(m) &= cm \log(m+1) \quad \text{if } G \text{ supports FFT,} \\ M(m) &= cm \log(m+1) \log \log(m+2) \quad \text{for any field.} \end{aligned}$$

The general bound with the extra $\log \log$ factor comes from an FFT method recursively applied to the polynomial rings $F[t]/(t^K + 1)$ for various values of $K = 2^k$, analogous to the corresponding fast integer multiplication technique. This approach fails for fields of characteristic 2, but then one can use

$$F[t]/(t^{2K} + t^K + 1)$$

with $K = 3^k$ instead (see [34]). Also for these algebraic models, no nonlinear lower bounds are known for the discrete Fourier transform or for polynomial multiplication.

The best upper bounds presently available for the complexity of solving the equation $a(t)x(t) = b(t) \bmod q(t)$ for an arbitrary m th degree polynomial $q(t)$ are higher by a factor of order $\log m$ again. Rather precise upper and lower bounds for the nonscalar complexity of extended polynomial gcd computations have been obtained by Strassen (see [49], where also further references can be found). A numerical version of the gcd problem (over the field \mathbf{C} , or \mathbf{R}) is treated in [40].

This last remark leads us to the crucial question to be studied next: What are the implications of the *algebraic* complexity results for the (machine bounded) time complexity of the corresponding *numerical* computations with polynomials?

By inserting numbers (elements of \mathbf{Z} , \mathbf{R} , or \mathbf{C}) for the coefficients of the polynomials, fast algebraic algorithms clearly should get transformed into fast numerical procedures (provided one has numerical stability), but there are also other kinds of *machine algorithms*, not obtainable in this way, and we cannot exclude the possibility that some of these are significantly faster than anything constructed purely algebraically. In fact, the numerical multiplication of polynomials seems to furnish an example of this phenomenon. Simply combining the unit cost FFT bounds from (2.4) with the estimates (2.3) for integer multiplication will at best yield the time bound $O(m \log m \mu(N))$ —but one can do better!

Replacing the variable t by 2^K with some $K \geq 2N + \log(m+1) + 1$ reduces the multiplication of two m th degree polynomials with coefficients of at most N bits in length to one long integer multiplication of size $O(m(N + \log(m+1)))$, and for $N > \log(m+1)$ that leads to the improved time bound $O(\mu(mN))$. For pointer machines we thus have even the linear bound $O(mN)$. The linearity in m also shows that, at least asymptotically, the celebrated FFT is not an optimal numerical method for the discrete Fourier transform $\mathbf{C}^m \rightarrow \mathbf{C}^m$. The details of this approach can be found in [37].

When multiplying polynomials with complex coefficients in this way via fast integer multiplication modulo $2^{2L} + 1$ one can nicely exploit that $2^L = \text{sqrt}(-1)$

can serve as the imaginary unit in this domain. This feature considerably enhances the applicability of the method.

Numerical *division* of polynomials requires some extra care with respect to stability. For the division by some (complex) polynomial F , its leading coefficient should be bounded away from zero. A good way to express this quantitatively is by an upper bound on the *root radius* $\rho(F)$ (defined as the maximal modulus of the roots of F in \mathbb{C}), together with some normalization of the *size* of F , measured by the l^1 -norm $|F|$ of the coefficient vector. The corresponding result from [37] is

(2.5) Numerical division of $G \in \Pi_m$ by $F \in \Pi_n$ within error 2^{-N} , i.e., computing some $Q \in \Pi_{m-n}$ and $R \in \Pi_{n-1}$ such that $|G - QF - R| < 2^{-N}$, is possible in pointer time $O(m(N + m + m \log(1 + r)))$, where r is a *given* bound on $\rho(F)$, and $|G| \leq 1 \leq |F| \leq 2$ is assumed.

Restricted to divisions with a uniform bound on the root radius of the denominators, the given bound on pointer time is of order $O(mN)$ again, provided the precision N is of order m at least.

3. The fundamental theorem of algebra in terms of computational complexity. For many centuries the problem of solving polynomial equations was mainly studied in terms of *formulae*. After the early days when Renaissance mathematicians had mastered the third and fourth degree it took more than 250 years until Gauss gave rigorous proofs for the *existence* of the roots of any real equation and Abel could show that, in general, one cannot obtain them by merely applying arithmetical operations and successively solving pure equations. This narrow notion of ‘solvability’ has been carried on up to this day, at least verbally, despite the rather clear comments in §9 of Gauss’s dissertation that ‘*resolutio aequationis*’ and ‘*ipsius reductio ad aequationes puras*’ should properly be distinguished.

It was only after about another one hundred years that solving general equations with complex coefficients was understood adequately, namely in terms of *algorithms*. Weyl’s constructive proof [53] for the fundamental theorem of algebra essentially shows that the zeros of a complex polynomial depend on its coefficients *recursively* (see also Specker’s contribution in [11]). In other words, there exists a Turing machine which, when fed with an integer n and with oracles for the coefficients of some m th degree polynomial, will output approximations for the m zeros of this polynomial within an error bound of 2^{-n} . In terms of *computational complexity* the decisive question is now: *how fast* can this be done?

3.1. The computational problem. It is a well-known fact that clustered zeros of a polynomial are less stable under small perturbations of the coefficients than isolated zeros. Therefore it seems to be more appropriate to understand any prescribed accuracy in the sense of backward analysis: a collection of approximate zeros should be considered acceptable, if their elementary symmetric functions agree well enough with the coefficients of the given polynomial. This applies to $P \in \Pi_m$, $P(z) = a_0 + a_1z + \cdots + a_mz^m$ with leading coefficient $a_m = 1$, but

more generally also $a_m \rightarrow 0$ should be admitted, i.e., some of the zeros may tend to infinity. In this way we are led to the following problem specification.

(3.1) Computational task of *approximate factorization*: Given any integer $N > 0$ and a polynomial $P \in \Pi_m$ of norm $|P| \leq 1$, compute linear factors $L_j(z) = u_j z + v_j$ ($1 \leq j \leq m$) such that $|P - L_1 L_2 \cdots L_m| < 2^{-N}$ is satisfied.

Here and in the sequel $|\cdot|$ stands for the l^1 -norm of the coefficient vectors. Using other common norms like the l^2 -norm or $\max_{|z|=1} |p(z)|$ would not make much of a difference, as the trade-off factors are less than $m+1$, which means a variation of N by not more than $\log(m+1)$.

In analyzing the efficiency of algorithms for this task we are mainly interested in their *worst case behavior* (note that the ‘worst’ case need not be really ‘bad’). For any such algorithm and for fixed values of the parameters m, N , there is a well-defined maximal running time $T(m, N)$, maximal with respect to all inputs (oracles) (a_0, \dots, a_m) of l^1 -norm ≤ 1 . (If this maximum would not exist, then König’s lemma would imply the existence of certain inputs for which the algorithm would never stop.) Now we can restate the main problem of this section more precisely: What is the true order of growth of the maximal running time $T(m, N)$ of (nearly) optimal algorithms for approximate factorization, what is the asymptotic behavior of $T(m, N)$ for $N \rightarrow \infty$, or for increasing degree? Most likely the answers will depend on the underlying machine model to a certain extent. In the light of our incomplete knowledge about much simpler complexity problems like integer multiplication we cannot hope for more than partial answers here. Our main result, presented below, is a good upper bound for the complexity of approximate factorization, but before going into this, we shall briefly review some of the partial answers which, at least implicitly, are contained in the vast literature on root finding methods.

In 1967 a symposium on constructive aspects of the fundamental theorem of algebra was held; the proceedings [11] provide a good record of the state of the art at that time. Furthermore, Chapter 6 of Henrici’s book [14] gives an extensive account of existing techniques. Many of the numerical methods used in practice try to compromise between speed and universal applicability. A major difficulty arises from the fact that all reliable algorithms for approximate factorization must necessarily employ multiprecision techniques which are usually slow, and rigorous a priori estimates for the round-off errors are required in order to guarantee reliable results. The same applies to all iterative methods usually posing the additional difficulty of finding suitable starting values.

By adapting the ideas found in the numerical literature to the strict conventions explained before we arrive at the conclusion that approximate factorization is certainly possible in *polynomial time*. On the base of Weyl’s exclusion method as worked out by Henrici [14, pp. 517–522] we can, for instance, obtain the estimate $T(m, N) = O(m^7 N^3)$. Similar time bounds appear for methods of root isolation developed in Computer Algebra (see [8] and the references given there). There remains the challenge of finding more precise bounds. Application of the

fast multiplication techniques for integers and polynomials (cf. 2.1 and 2.2) reduces the bound mentioned to something like $O(m^{5+\delta}N^{2+\delta})$ for any $\delta > 0$, but the factor N^2 seems to be unavoidable with these methods.

3.2. An upper bound. The main result to be presented here is a good upper bound on the time complexity of approximate factorization found by the author in 1981/82. It is based upon a new technique called the *splitting circle method* ("Trennkreisverfahren" in German), described in the Preliminary Report [38]. Some of the material is highly technical, and many details of its implementation need to be worked out even more thoroughly. A full account of the new results will come out as a monograph. In the sequel we shall give a brief outline of the main ideas and results.

THEOREM. *Approximate factorization of complex polynomials as specified in (3.1) is possible within maximal running time*

$$T(m, N) = O(m\mu(m^2 \log m) + m\mu(mN)), \quad (3.2)$$

with respect to the machine models and the corresponding time bounds on integer multiplication listed in (2.3).

In order to simplify the presentation we shall restrict further discussion to the case of pointer time with the linear bound $\mu(N) = cN$. (In case of the other models additional logarithmic factors come in.) Then the upper bound becomes $T(m, N) = O(m^3 \log n + m^2 N)$. It will be instructive to compare this with obvious time bounds for the corresponding verification problem. In doing so we may assume, as a general rule of thumb, that computations with polynomials of degree of order m should be carried out with m -bit precision at least, because of the following theoretical observations.

First of all, one should be aware of the fact that small polynomials may have big factors. For $m = 2k$ the simple example $P(z) = 2^{-k}(z^2 - i)^k = 2^{-k}(z - a)^k(z + a)^k$ (with $a^2 = i$) shows a partial factorization $P = FG$ with $|P| = |F| = 1$, but $|G| = 2^{m/2}$, thus the corresponding amplification of errors previously made in the computation of F will necessitate an extra precision of $m/2$ bits. The precise upper bound in the case of many factors is contained in the following *quantitative supplement to the fundamental theorem of algebra*—usually not found in textbooks!

$$\text{Let } f \in \Pi_m, \quad f = f_1 \cdots f_k; \quad \text{then } |f| \leq |f_1| \cdots |f_k| \leq 2^{m-1}|f|. \quad (3.3)$$

The upper bound is attained for such tame polynomials as $f(z) = z^m - 1$ in case of its complete factorization, i.e., for $k = m$.

Similar error amplification may be caused by *Taylor shifts* (for fixed a the linear mapping defined on Π_m by $g(z) = f(z + a)$ has operator norm $(1 + |a|)^m$), or by the numerical division of polynomials, see (2.5). Assuming $N \geq m$ we thus see that any naive method just for *checking* the accuracy of a given approximate factorization will require $O(m^4)$ bit operations at least, and even by means of fast multiplication techniques we cannot do better than in pointer time of

order $Nm \log m$, which in fact seems to be the precise order of growth for the complexity of this verification problem.

Therefore the bound (3.2) seems to be optimal up to a factor of order m at most. For increasing accuracy the second term $O(m^2 N)$ becomes dominant, and for a wide class of polynomials admitting *balanced splittings* (explained below) it can be replaced by the smaller bound $O(Nm \log m)$. The main advantage with all these bounds is their linearity in N . For any *fixed* degree m the bound (3.2) becomes simply $O(\mu(N))$ (with the implicit constants depending on m), and this is indeed optimal. For pointer machines we arrive at a linear time bound, not more than a constant multiple of the time needed to output the results. But even in the case of machine models, for which the precise complexity of integer multiplication is yet unknown, we have the following relative result.

(3.4) THEOREM. *For $N \rightarrow \infty$ and fixed $k, m \geq 2$ the complexity bounds for the following tasks all have the same order of growth:*

- (a) *N -bit integer multiplication,*
- (b) *computing $(1+x)^{1/k}$ for $|x| \leq 1/2$ within error bound 2^{-N} ,*
- (c) *approximate factorization of m th degree polynomials within error 2^{-N} .*

For bounding (c) by (a) again some regularity condition is assumed, as needed for (2.2), and then (3.2) applies. Approximate factorization of $z^k - (1+x)$ reduces (b) to (c) for $k \leq m$, and in general via (a). The transition from (b) to (a) is by an idea due to H. Alt, to simulate integer squaring by means of a sufficiently precise evaluation of

$$(2 - (1+x)^{1/k} - (1-x)^{1/k})k^2/(k-1) = x^2 + O(x^4)$$

for small x .

In §3.4 we will provide the perturbation argument needed to infer from an approximate factorization to approximate *solutions* (the zeros) of a polynomial equation, with leading coefficient one and bounded coefficients, say. In case of fixed degree this will amplify the time bounds by not more than another constant factor. Therefore the equivalence of (b) and (c) shows that, in contrast to Abel's and Galois's findings on "solubility," in terms of computational complexity there is no significant difference between solving pure or general equations with complex coefficients. Computationally, the approximate solution of general polynomial equations is reduced to the obviously more fundamental task of integer multiplication.

3.3. *The splitting circle method.* The general strategy for the approximate factorization of a given $P \in \Pi_m$ is based on a substrategy for approximately splitting P into two factors $F \in \Pi_k$, $G \in \Pi_{m-k}$, to be applied recursively. The case $k = 1$ simply means approximate determination of a single zero and its deflation. High accuracy can be achieved fast by means of Newton iteration, provided the zero is simple, well isolated, and a suitable starting value is known. Similarly the case $k = 2$ is covered by Bairstow's method, quadratically convergent, if the two zeros are well separated from the $m - 2$ other zeros. The

splitting circle method employs the corresponding general Newton iteration (cf. J. Schröder's paper in [11]), where the choice of a suitable k will depend on the distribution of the zeros, such that there are k zeros of P *inside* of some suitable *splitting circle* while the $m - k$ other zeros lie *outside* of this circle, possibly some of them close to infinity. Moreover all the zeros shall stay at some distance of the circle. It is a fundamental fact that such a circle can always be determined, except for the singular cases $P(z) \approx (az + b)^m$ which will be found out properly to be approximately factored already.

There are stable transformations (scalings, Taylor shifts), by which any circle can be reduced to the standard case of the unit circle $E = \{z: |z| = 1\}$. Due to the prior condition of a zero free annulus around E it is possible to obtain a reasonable lower bound on the decisive quantity $\nu = \min_{|z|=1} |P(z)|$, both theoretically and computationally with a rather moderate amount of work. The details in choosing such splitting circles can be arranged such that always $\log(1/\nu) \leq O(m)$. Based on this ν the analysis of the general Newton method then leads to explicit bounds for its quadratic convergence in the following way.

Given some approximate unit circle splitting $|P - FG| < \varepsilon$ (i.e., the k roots of F and the reciprocals of the $m - k$ roots of G are bounded by $1 - \delta$), the *Newton correction* is a pair $(f, g) \in \Pi_{k-1} \oplus \Pi_{m-k-1}$ to be chosen such that terms up to first order will cancel each other. Hence the new approximate factors $F + f$ and $G + g$ will satisfy

$$P - (F + f)(G + g) = P - FG - fG - gF - fg = -fg$$

with the new error being bounded by $|fg| \leq |f||g|$, where f and g are uniquely determined by the (incomplete) partial fraction decomposition $(P - FG)/(FG) = f/F + g/G$. Thus one can use the integral representation

$$f(z) = \frac{1}{2\pi i} \int_E \frac{(P - FG)(t)}{(FG)(t)} \frac{F(z) - F(t)}{z - t} dt$$

for deriving an upper bound for $|f|$, and similarly for $|g|$. In this way one shows, in particular, that an initial approximate splitting with precision

$$|P - F_0 G_0| < \varepsilon_0 = \nu^4 2^{-cm} \quad \text{with some constant } c$$

will suffice.

Finding such initial F_0 and G_0 is possible by means of contour integration (also used in [12], though without error bounds). The power sums s_j of the k zeros of P which lie inside of E are expressible as

$$s_j = \frac{1}{2\pi i} \int_E \frac{P'(z)}{P(z)} z^j dz.$$

Here the zero free annulus and the lower bound ν enable us to give explicit bounds for the precision required for the simultaneous evaluation of these integrals by means of a discrete Fourier transform. (This step essentially contributes to the first term of the upper bound (3.2).) The approximations for $s_0 = k$, s_1, \dots, s_k

easily yield an approximate factor F_0 , and a suitable G_0 can then be found by polynomial division.

So far we have tacitly assumed that the zeros of the given polynomial P can (in advance) be localized well enough to admit a proper choice of the splitting circles, but this is actually the key problem of the whole method. Our algorithms for the corresponding *diagnostics* are based on the classical Graeffe process of *root squaring*. We use a variant especially suited to exploiting fast integer multiplication. Contrary to the common application of Graeffe's method, here the primary interest is not in circles *on* which the zeros are located but in circles bounded away from all the zeros, and for the latter purpose moderate precision is quite sufficient. Moreover one should not try to compute all the root moduli $\rho_1(f) \geq \rho_2(f) \geq \dots \geq \rho_m(f)$ of some f simultaneously, which indeed may require prohibitive (i.e., too expensive) long number computations (cf. the discussion in Turán's paper [52]). Instead, one can shift the *focus of precision* by suitable scalings after each root squaring step such that the algorithm is directed to the computation of one particular $\rho_k(f)$. The following time bound is a typical result obtained in this way.

(3.5) Given any complex polynomial $f(z) = a_0 + a_1z + \dots + a_mz^m$ with $|f| \leq 1$, given some positive $\delta < 1/2$ and an index $k \leq m$, computing the k th root modulus $\rho_k(f)$ with relative error less than δ is possible in pointer time $O(m^2(\log m + \log(1/\delta))(\log \log m + \log(1/\delta)))$. (In addition, something like $|a_0|, |a_m| \geq (\delta/m)^m$ should hold.)

Applied with $\delta = m^{-O(1)}$ this leads to the rather favorable time bound $O(m^2 \log^2 m)$, while this approach cannot be recommended for much higher precision requirements, where the quadratic growth in $\log(1/\delta)$ becomes dominant.

For other similar estimates and further details we must refer to [38]. The overall time analysis of the splitting circle method reveals that balanced splittings (with k and $m - k$ of the same size, ideally $k = m/2$) are preferable. Even in the case of m well isolated simple zeros the deflation of one linear factor after the other is inferior compared with the *divide and conquer* principle of computational complexity, but there do exist certain hard polynomials which, by their particular zero spacing, will enforce very unbalanced splittings. A whole family of such examples (with arbitrary parameters $1 < u_j < 1.01$) is furnished by

$$f(z) = (z + u_1/4)(z + u_2/16)(z + u_3/64) \cdots (z + u_m/4^m).$$

Here $N = 2m^2$ specifies a precision well tuned with respect to the coefficient size. This property of being *strongly nested* represents a new phenomenon in the numerical treatment of polynomials. At present it is hard to say whether such polynomials are really ill-conditioned with regard to the task of approximate factorization, or whether this just shows a particular weakness of the splitting circle method. In any case this nestedness should properly be distinguished from the occurrence of clustered zeros which, in itself, does not cause extra difficulties for the approximate factorization of a polynomial, thanks to a blow-up technique

by proper scaling (see [38]). The presence of clusters will, however, affect the *stability* of the zeros, to be discussed below.

3.4. Some applications. Any approximate factorization $|P - L_1 \cdots L_m| < \varepsilon$ with $0.5 \leq |P| \leq 1$ and known linear factors $L_j(z) = u_j z + v_j$ can be used for the approximate determination of the zeros of P . The natural candidates for this approximation are the numbers $w_j = -v_j/u_j$; in case of $|v_j| > |u_j|$ it may be preferable to consider their reciprocals as approximations for the reciprocals of the zeros of P .

The corresponding inequalities $|P(w_j)| < \varepsilon$ (or $|P^*(1/w_j)| < \varepsilon$ with the reversed polynomial $P^*(z) = z^m P(1/z)$) combined with a well-known homotopy argument induce $W = \{z: |P(z)| < \varepsilon \max(1, |z|^m)\}$ as an *inclusion set* for the zeros, and their perturbations are bounded by the diameters of the components of W . Here we want to consider (worst case) a priori bounds only, while in practical applications one should, of course, take advantage of possible shortcuts based on better a posteriori bounds. Keeping to the previous notations we have the following perturbation bound (Theorem 2.7 in [40]).

(3.6) The zeros z_1, \dots, z_m of P can be numbered such that (for $\varepsilon < 2^{-7m}$)

$$|z_j - w_j| < 9\varepsilon^{1/m} \quad \text{for } |w_j| \leq 1, \quad |1/z_j - 1/w_j| < 9\varepsilon^{1/m} \quad \text{for } |w_j| \geq 1.$$

This is optimal up to the constant factor, and is better by a factor of order m than corresponding bounds given by Ostrowski (cf. [29, Appendices A, B]). Aiming at an error bound of 2^{-n} for the roots of some m th degree polynomial we see that approximate factorization with $\varepsilon = 2^{-N}$, $N = (n+4)m$ will always suffice, whence the time bound (3.2) yields

(3.7) Computing the zeros (or their reciprocals) of any m th degree polynomial P (of norm $0.5 \leq |P| \leq 1$, say) with error less than 2^{-n} is possible in pointer time $O(m^3 \log m + m^3 n)$.

There are applications, where the determination of just one zero will suffice. For that purpose the splitting circle method can be modified such that the second term in the time bound of (3.7) may be replaced by the smaller bound $O(m^2 n)$, since after a first approximate splitting of P into two factors F and G , only one of these (of degree $\leq m/2$) needs further treatment, etc., recursively. Accordingly, in this case unbalanced splittings are most welcome.

As an example of this kind we like to mention a rather fast method for the factorization of univariate integer polynomials based on high precision computation of one zero and on diophantine approximation to find its minimal polynomial as a factor (see [26] and [39]). Along these lines extensive numerical calculations with *algebraic numbers* seem to become feasible.

Another important consequence of the above time bounds is that *all algebraic functions can be computed fast*. In their paper [20] with exactly this title, Kung and Traub investigate the complexity of the somewhat different task to compute an initial segment of one of the (fractional) power series expansions of an algebraic function, counting arithmetical operations at unit cost, where finding

a zero of a numerical polynomial (to any prescribed precision) is regarded as a primitive operation.

In our model this task cannot be solvable in general, since testing for equality is impossible. Let us consider a more direct approach, instead. Given any polynomial $F(z, v) = a_0(v) + a_1(v)z + \cdots + a_m(v)z^m$ and some point $v = (v_1, \dots, v_k)$ i.e., the coefficients of the polynomials $a_j \in \mathbb{C}[v_1, \dots, v_k]$ and the point v are available from oracles again, the m branches of the algebraic function $z(v)$ defined by $F(z, v) = 0$ can be evaluated at v within time bounds of the same order as in (3.7), provided some normalizing assumptions about v and the a 's are made. The major strength of this assertion lies in its *uniformity* with respect to v ; the hidden constants can be chosen, for instance, to be valid for all $|v| \leq 1$.

4. Linear equations and the complexity of matrix computations.

One of the pioneering results in early complexity theory was Strassen's discovery [46] that "Gaussian elimination is not optimal." He showed how to replace the classical $O(n^3)$ methods of matrix multiplication and inversion by algorithms which require less than $O(n^{2.81})$ arithmetical operations. The basic idea is quite elementary and widely known nowadays, though most numerical analysts have remained sceptical about the practical applicability of such methods.

Meanwhile the exponent 2.81 has been improved substantially by the efforts of several authors (cf. [3, 36, 10] and Pan's survey [30]). The best bound presently known to the author is a bit smaller than 2.4785, presented by V. Strassen at an informal meeting in 1985 (see also [51]). Actually these bounds concern the *complexity of matrix multiplication*. In §§4.1 and 4.2 we will briefly review some of the main ideas and results related to this fascinating unsolved mathematical problem; possibly this outline could stimulate other mathematicians to contribute to its complete solution.

In §4.3 we will discuss the complexity of solving linear systems and of other matrix computations like matrix inversion, evaluating determinants, etc., mainly in the framework of algebraic complexity, where the elements of matrices and vectors are considered as indeterminates over some ground field F . When dealing with real or complex numbers, we have to include numerical aspects again, especially for the complexity bounds on the approximate determination of eigenvalues.

With regard to other domains we should mention here that linear systems of diophantine equations can be solved in polynomial time as shown by J. von zur Gathen and M. Sieveking (in [45, pp. 57–65]). Related algorithms and further references are found in [7] and [16].

4.1. The exponent of matrix multiplication. The algebraic complexity of matrix multiplication can be measured in different ways. We consider $n \times n$ matrices $A = (a \dots)$, $B = (b \dots)$ with indeterminate entries over some scalar field F (to be held fixed for the moment). Let $L(n)$ denote the minimal length of straight-line programs in $F(a \dots, b \dots)$ computing the entries of AB , where all arithmetical

operations are counted, and let $L^{\wedge}(n)$ be the corresponding nonscalar complexity, or $L^*(n)$ in case of $F[a \dots, b \dots]$, if divisions shall not be used. For infinite F we have $L^{\wedge}(n) = L^*(n)$ (cf. [47]).

Last but not least, the *bilinear complexity* $L^{\otimes}(n)$ refers to the model where only multiplications $\xi(a \dots) * \eta(b \dots)$ of F -linear forms in the a 's by F -linear forms in the b 's are admitted.

The *exponent of matrix multiplication* (over F) is defined as the infimum

$$\omega(F) = \inf\{\beta: L(n) = O(n^{\beta})\}, \quad (4.1)$$

obviously satisfying $2 \leq \omega(F) \leq 3$. Strassen's first nontrivial bound $\omega(F) \leq \log 7 / \log 2 < 2.81$ was derived from $L^{\otimes}(2) \leq 7$. In the same way any bilinear algorithm for some other small n can be applied to multiply $n \times n$ *block matrices* recursively; thus $L^{\otimes}(n) \leq r$ implies $L(N) = O(N^{\beta})$ with $\beta = \log r / \log n$. Combined with the elementary inequalities

$$L^{\wedge}(n) \leq L^*(n) \leq L^{\otimes}(n) \leq 2L^*(n),$$

this shows that we may replace $L(n)$ in (4.1) by any of these other measures without altering the definition of $\omega(F)$. The further discussion will concentrate on $L^{\otimes}(n)$ and its generalization to the multiplication of rectangular matrices. For technical details and proofs see [36].

Let $\langle k, m, n \rangle$ denote the *tensor* for the bilinear mapping of multiplying $k \times m$ - with $m \times n$ -matrices. Its *rank* $\text{rk}\langle k, m, n \rangle$ is equal to the minimal length of representations of the associated trilinear form $\text{trace}(ABC)$ as a sum of *rank one products* of linear forms, where C has format $n \times k$. Spelled out in coordinates this means, with $A = (a \dots)$, $B = (b \dots)$, $C = (c \dots)$, that $\text{rk}\langle k, m, n \rangle \leq r$ is equivalent to the existence of linear forms ξ_j in the a 's, η_j in the b 's, ζ_j in the c 's such that

$$\sum_{j=1}^r \xi_j(a \dots) \eta_j(b \dots) \zeta_j(c \dots) = \text{tr}(ABC) = \sum_{\kappa, \mu, \nu} a_{\kappa, \mu} b_{\mu, \nu} c_{\nu, \kappa} \quad (4.2)$$

is satisfied, which in turn is equivalent to the solvability of a huge system of $k^2 m^2 n^2$ nonlinear equations over F for the $kmr + mnr + nkr$ unknown coefficients of the linear forms. The formidable size of such systems may give an impression of the difficulties to determine any of the values of $\text{rk}\langle k, m, n \rangle$, even in the case of low dimensions (despite the fact that, for fields like \mathbf{R} or \mathbf{C} with decidable first order theory, $\text{rk}\langle k, m, n \rangle$ is a *computable* function). One of the rare values explicitly known is $\text{rk}\langle 2, 2, 2 \rangle = 7$ for any field F , as always $\text{rk}\langle n, n, n \rangle = L^{\otimes}(n) \geq L^*(n) \geq 2n^2 - 1$; the latter inequality is an instance of a more general lower bound on the complexity of associative algebras (see [1, 50]). The next higher case of 3×3 matrices would already amount to a system of 729 equations with 567 unknowns (for $r = 21$); so far nothing better than $\text{rk}\langle 3, 3, 3 \rangle \leq 23$ is known. More generally, one could also hope to derive better estimates for $\omega(F)$ from good upper rank bounds for *rectangular* formats, by means of the following lemma.

$$\begin{aligned} \text{rk}\langle k, m, n \rangle \text{ is symmetric in } k, m, n, \text{ and submultiplicative, i.e.,} \\ \text{rk}(\langle k', m', n' \rangle \otimes \langle k'', m'', n'' \rangle) \leq \text{rk}\langle k', m', n' \rangle \text{rk}\langle k'', m'', n'' \rangle. \end{aligned} \quad (4.3)$$

Starting from some bound $\text{rk}\langle k, m, n \rangle \leq r$ we apply symmetrization and obtain $L^{\otimes}(kmn) = \text{rk}\langle kmn, mnk, nkm \rangle \leq r^3$; therefore,

$$\text{rk}\langle k, m, n \rangle \leq r = (kmn)^{\tau} \quad \text{implies} \quad \omega(F) \leq 3\tau. \quad (4.4)$$

So far, however, none of the rank bounds available for *small* formats has proved to be good enough to yield better estimates than 2.81. In 1978, this barrier was broken by V. Pan. By means of his technique of *trilinear aggregating* he could show $L^*(n) \leq n^3/3 + O(n^2)$ with such constants that $n = 48$ led to the improved estimate $\omega(F) < 2.781$.

A similar bound was obtained by Bini, Capovani, Lotti, Romani [3], but they used a completely new approach based on the notion of *border rank*. Let ε be a further indeterminate over F . The border rank $\text{brk}\langle k, m, n \rangle$ can be defined as the minimal length of sum representations (4.2) up to errors $O(\varepsilon)$, where now *meromorphic formal power series* in ε are admitted as coefficients of the linear forms. (For a definition in the language of algebraic geometry see §§13, 14 of [50].)

The properties (4.3) and (4.4) are also true for border rank, whence

$$\text{brk}\langle k, m, n \rangle \leq r = (kmn)^{\tau} \quad \text{implies} \quad \omega(F) \leq 3\tau. \quad (4.4b)$$

In [3] this was applied to $\text{brk}\langle 2, 2, 3 \rangle \leq 10$; another example is $\text{brk}\langle 3, 3, 3 \rangle \leq 21$, which yields $\omega(F) < 2.772$. For the paradigmatic case of 2×2 matrices one knows the bounds $6 \leq \text{brk}\langle 2, 2, 2 \rangle \leq 7$.

The proof of (4.4b) exploits the fact that the tensorial powers of any tensor t with $\text{brk}t \leq r$ satisfy a rank estimate $\text{rk}(t^{\otimes s}) \leq O(s^2) \cdot r^s$; the extra factor of polynomial growth does not matter when (4.4) is applied for $s \rightarrow \infty$. Similar arguments are used to show $\omega(F) = \omega(F_0)$, where F_0 denotes the prime field of F (see [36, Theorem 2.8]), thus the exponent of matrix multiplication over F can only depend on the characteristic of F . In the sequel we will simply write ω , since all further discussions are uniform with respect to F .

4.2. Subadditivity. There exist computational problems which admit substantial savings under *mass production*. An important example is furnished by the evaluation of a general m th degree polynomial at m different points (cf. [4]). Similar effects occur in the realm of matrix computations. Multiplying one column vector b by an $n \times n$ matrix A takes exactly n^2 multiplications, while multiplying n different vectors b_1, \dots, b_n by the same A is possible by one matrix multiplication—in less than $O(n^{2.8})$ steps. One could argue that these examples are relying on a certain amount of joint information (the coefficients of the polynomial, the matrix A). So let us impose very strict rules now. We consider two completely *disjoint problems* of matrix multiplication and ask whether forming AB and UV in one compound computation may be cheaper than by performing (optimal) algorithms for the two matrix multiplications separately.

Under the bilinear cost measure this is equivalent to the question whether the *subadditivity* of rk , i.e. (with respect to direct sums),

$$\text{rk}(\langle k', m', n' \rangle \oplus \langle k'', m'', n'' \rangle) \leq \text{rk}\langle k', m', n' \rangle + \text{rk}\langle k'', m'', n'' \rangle, \quad (4.5)$$

can become a strict inequality. For 2-stage tensors (matrices) equality always holds, while the corresponding *additivity conjecture* (cf. [47]) for 3-stage tensors is still an open problem.

Subadditivity is true for border rank as well, but in that case we definitely know that additivity does not hold in general. Already for the special class of tensors $\langle k, m, n \rangle$, the left-hand and the right-hand side of (4.5) (with brk instead of rk) can differ unboundedly. We present an example from [36] which has played an important role in estimating the exponent of matrix multiplication. Based on a dimension argument, one has $brk\langle 3, 1, 3 \rangle = 9$ and $brk\langle 1, 4, 1 \rangle = 4$, but also

$$brk(\langle 3, 1, 3 \rangle \oplus \langle 1, 4, 1 \rangle) = 10. \quad (4.6)$$

We show that the trilinear form $\text{tr}(ABC) + \text{tr}(UVW)$ associated with the direct sum of these disjoint problems indeed has an approximate decomposition (4.2) of length 10, with coefficients from $F((\varepsilon))$ and up to an error term $O(\varepsilon)$, namely

$$\begin{aligned} & \sum_{i=1}^3 \sum_{j=1}^3 (a_i + \varepsilon u_{i,j})(b_j + \varepsilon v_{i,j})(c_{j,i} + \varepsilon^{-2}w) - \left(\sum_i a_i \right) \left(\sum_j b_j \right) \varepsilon^{-2}w \\ &= \sum_{i=1}^3 \sum_{j=1}^3 (a_i b_j c_{j,i} + u_{i,j} v_{i,j} w) + O(\varepsilon), \end{aligned}$$

where the u 's and v 's with subscripts $(i, j) = (1, 1), (1, 2), (2, 1), (2, 2)$ are the indeterminate components of two vectors of length 4, while the remaining u 's and v 's are chosen to yield proper canceling, namely

$$u_{i,3} = v_{3,j} = 0, \quad u_{3,j} = -u_{1,j} - u_{2,j}, \quad v_{i,3} = -v_{i,1} - v_{i,2}.$$

Similar to Strassen's $rk\langle 2, 2, 2 \rangle \leq 7$ versus the trivial number of 8 multiplications, which gave the estimate $\omega \leq 3 \log 7 / \log 8$, we can regard (4.6) as a bound of 10 on the border rank of a problem of *partial* matrix multiplication with the trivial number of 13 multiplications, and this indeed implies $\omega \leq 3 \log 10 / \log 13 < 2.7$, by a corresponding extension of (4.4b) [36, Theorem 4.1]. There is, however, an even more productive way of estimating ω on the base of such examples, which in addition exploits the *disjointness* of the pieces. This stronger generalization of (4.4b) is the following

τ -THEOREM (de Groote's naming for a special case of Theorem 7.1 from [36]).

$$brk(\langle k_1, m_1, n_1 \rangle \oplus \cdots \oplus \langle k_p, m_p, n_p \rangle) \leq r = \sum_{i=1}^p (k_i m_i n_i)^\tau \quad \text{implies} \quad \omega \leq 3\tau.$$

With regard to (4.6), for instance, we solve the equation $9^\tau + 4^\tau = 10$ and obtain the better bound $\omega < 2.6$.

The presentation of this theorem at an Oberwolfach conference in 1979 initiated a hot competition among specialists to find better and better estimates from more and more sophisticated designs, similar to (4.6), and the bounds for ω

came closer and closer to 2.5. Corresponding lower bound speculations, however, were soon abandoned by Coppersmith and Winograd (see [10]). They found a general method for the iterated construction of such examples, and by starting from (4.6) they succeeded in showing $\omega < 2.5$. Moreover, it is an important theoretical consequence of their work that any particular instance of the τ -theorem admits some further improvement yielding a slightly better bound for ω , i.e., we always have the strict inequality $\omega < 3\tau$.

Strassen's new method [51] stems from a thorough investigation of rank and border rank in a more general algebraic setting. His constructions have led to an application of the τ -theorem combined with a limit process which yields $\omega \leq \log((h+1)^3/4)/\log h$ (for $h = 2, 3, \dots$). The best estimate is obtained for $h = 5$, namely $\omega \leq \log 54/\log 5 < 2.4785$.

4.3. *Linear systems and matrix inversion.* Solving a general system $Ax = b$ of n linear equations in n unknowns x_1, \dots, x_n can be considered as the task of computing the x 's as rational functions in the indeterminate entries of A and b . Optimal algorithms for this task are straight-line programs of minimal length in $F(a \dots, b \dots)$, and it was with respect to this model that Gaussian elimination was originally shown not to be optimal. (For $n = 2$, by the way, and under the measure of nonscalar complexity, Gaussian elimination *is* in fact optimal, see [28].) The solution of the system can simply be expressed as $x = A^{-1}b$. Thus solving such a general system is possible within the upper bound $I(n) + 2n^2 - n$, where $I(n)$ denotes the minimal number of arithmetical operations sufficient for the computation of the inverse of a general $n \times n$ matrix. In [46], 2×2 *block inversion* was applied recursively to reduce matrix inversion to (fast) matrix multiplication. The corresponding estimate is

$$I(2n) \leq 2I(n) + 6L(n) + n^2 + n.$$

Similar reasoning shows that the complexity $D(n)$ of computing $n \times n$ determinants satisfies the recursive inequality

$$D(2n) \leq I(n) + 2L(n) + 2D(n) + n^2 + 1.$$

Therefore $L(n) = O(n^\beta)$ (for any $\beta > \omega$) implies that $I(n)$ and $D(n)$ are of the same order of growth at most. Conversely we can also show that the complexity of matrix inversion is not much smaller than that of matrix multiplication. With such proofs, however, one should keep in mind that straight-line programs valid for the model with indeterminates will usually not work for all specializations by nonsingular matrices, since substitution of algebraically dependent elements may cause divisions by zero. Below we will discuss other models without this deficiency.

The identity $X^2 = (X^{-1} - (I + X)^{-1})^{-1} - X$ shows that the complexity $Q(n)$ of *squaring* a general $n \times n$ matrix satisfies $Q(n) \leq 3I(n) + 2n^2 + n$. The multiplication of general $2n \times 2n$ matrices,

$$\begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} \begin{pmatrix} B_1 & B_3 \\ B_2 & B_4 \end{pmatrix}$$

with $n \times n$ blocks A_i, B_j is reduced to squarings in the following way. With $S = B_1 + B_2$, the first n columns of the result are obtained from

$$\begin{pmatrix} A_3 - S & B_1 \\ A_1 & S \end{pmatrix}^2 - \begin{pmatrix} A_4 - S & -B_2 \\ A_2 & S \end{pmatrix}^2 = \begin{pmatrix} \cdots & A_3B_1 + A_4B_2 \\ \cdots & A_1B_1 + A_2B_2 \end{pmatrix},$$

and equally for the second half. That shows $L(2n) \leq 4Q(2n) + 12n^2$, whence $L(2n) \leq 12I(2n) + O(n^2)$. For infinite fields F a similar estimate (with other constants) can be found in [47].

There is also a reduction from general matrix inversion to the computation of the *determinant*, obtained by Baur and Strassen as a corollary to their following general theorem (cf. [2]):

(4.7) If $f \in F(x_1, x_2, \dots, x_k)$ is computable by a straight-line program of length s , then the complexity of computing f together with all its partial first derivatives $\partial f / \partial x_j$ is bounded by $4s$.

Another illustrating example of these reduction techniques is the following quick proof for K. Kalorkoti's (unpublished) result that computing the *trace of the inverse* is as difficult as matrix multiplication, up to a constant factor (over infinite fields), hence essentially of the same complexity as computing the inverse itself. If there is a straight-line program for $\text{tr}(X^{-1})$ of length s , then $4s$ steps suffice for X^{-2} (built up from the partial derivatives); applying this once more we get $8s$ as an upper bound for the complexity of computing X^4 . After elimination of the divisions from such a program (at the cost of another constant factor, see [47]) we can finally specialize X as a suitable $3n \times 3n$ matrix such that computing X^4 will simulate a general matrix multiplication of size $n \times n$.

It is still an open question, whether the original problem of solving $n \times n$ systems, i.e., computing $A^{-1}b$ for general A, b , is powerful enough to simulate matrix multiplication or any of the other equivalent problems. Quite recently, however, T. Lickteig could show that the more general task of computing *all* solutions of *rectangular* systems $Ax = b$, which means, for instance, to compute a basis for the kernel of a general $n \times 2n$ matrix, has indeed the same order of complexity as matrix multiplication.

So far we have considered problems with algebraically independent inputs only, but with respect to applications other models seem to be preferable. In case of the prime fields with their finite extensions, *computation trees* with branchings based on zero testing furnish an adequate framework for the discussion of solving linear systems, computing the rank of matrices, or the inverse (if it exists), etc. Most of the reductions mentioned before similarly apply to this model. For further applications we refer to [17], including the proof that the algebraic complexity of computing the coefficients of the *characteristic polynomial* is of the same order as that of matrix multiplication, up to logarithmic factors.

With the possibility of zero testing, the open problem whether solving $Ax = b$ may have significantly lower complexity than matrix multiplication appears in a new light. The corresponding *verification* problem for a given solution x can obviously be solved within $O(n^2)$ steps. It is also an interesting question whether

verification of matrix multiplication, i.e., checking whether $AB = C$ for given C , could possibly be done faster.

The preceding model with zero testing does not apply to the fields \mathbf{R} or \mathbf{C} with oracle inputs (cf. §2) where equality is undecidable. It is, for instance, impossible to compute the rank of arbitrary matrices, or to decide whether a given $n \times n$ matrix A has determinant zero, but if we assume the regularity of A , then the fast algebraic methods of matrix multiplications can be utilized for the approximate computation of the inverse. By means of the hermitean transpose we have $A^{-1} = (A^H A)^{-1} A^H$, and for the inversion of positive definite matrices Strassen's block elimination (cf. [46]) is stable (provided the underlying method for matrix multiplication does not use divisions). The running time of such machine algorithms will, of course, also depend on the desired accuracy and on the condition of A .

When discussing methods for approximately solving $Ax = b$ one should properly distinguish whether some a priori bound on the condition of A is known or not. In the latter case it is rather likely that the major amount of work will be required to obtain sufficient information about the condition in order to guarantee a certain accuracy of the solution.

4.4. Characteristic equations. Finally we want to give a brief outline how to combine the time bounds on root finding from §3 with the preceding results on the algebraic complexity of matrix computations in order to obtain a good upper bound on the complexity of approximately computing the *eigenvalues* of arbitrary complex $m \times m$ matrices. This is just the very beginning of more extensive complexity studies of related problems like the determination of invariant subspaces, or improved time bounds for special classes of matrices, etc., still to be undertaken.

Let A be any $m \times m$ matrix of l^1 operator norm $|A| \leq 1$, or $\leq 1/2m$ after suitable scaling, and assume that the subsequent computations are done with N -bit precision. We describe a method for the computation of the coefficients of the characteristic polynomial of A which will mimic the algebraic approach from [17] numerically. Let ω denote the exponent of matrix multiplication over fields of characteristic zero, and let β be any number $\beta > \omega$. Then multiplication of complex $m \times m$ matrices in N -bit precision is possible in pointer time $O(m^\beta N)$. The first step in computing the characteristic polynomial is the approximate *unitary* transformation of A into *upper Hessenberg form* B ($b_{j+d,j} = 0$ for $d \geq 2$). By the methods from [32, 33], that is possible in pointer time $O(m^\beta N)$. Similar to the estimate (3.6), the perturbations of the eigenvalues caused by errors $|B - U^H A U| < \varepsilon$ are bounded by $O(\varepsilon^{1/m})$ in the worst case.

The next step is *scaling* by a similarity transformation with a diagonal matrix D such that $M = D^{-1} B D$ has only ones in its subdiagonal (without restriction we may assume that all the subdiagonal elements of B are nonzero; otherwise the problem would split into several smaller eigenvalue problems). Thus we get $M = S + R$, where R denotes the upper triangle of M , while the subdiagonal S represents a shift operator which maps the j th unit vector e_j upon e_{j+1} , with

$Se_m = 0$. The new entries

$$r_{k,k+d} = b_{k,k+d} \sum_{0 \leq j < d} b_{k+j+1,k+j} \quad (1 \leq k \leq k+d \leq m)$$

can be computed in a stable way by means of $O(m^2)$ multiplications of complex numbers bounded by one. Moreover we have $|R| \leq |B| \leq 1/2m$. The characteristic polynomial $f(z) = v_0 + v_1z + \dots + v_{m-1}z^{m-1} + z^m$ of M satisfies $f(M) = 0$. Applied to the first unit vector e_1 this equation yields $v_0x_1 + v_1x_2 + \dots + v_{m-1}x_m = -x_{m+1}$, where $x_j = M^{j-1}e_1$. Because of the dominance of S over $|R| \leq 1/2m$ one easily finds the estimate $|x_j - e_j| < e^{0.5} - 1 < 0.65$. Therefore the matrix X built up from the column vectors x_1, \dots, x_m satisfies $|X - I| < 0.65$, its inverse is bounded by $|X^{-1}| < 3$, whence the coefficient vector v can be obtained as $v = X^{-1}(-x_{m+1})$, numerically in pointer time $O(m^\beta N)$. The main idea from [17] concerns an efficient computation of the iterates x_j for $j \leq m+1$ by means of $O(\log m)$ matrix multiplications, first forming M^2, M^4, \dots , then $M^q(x_1, \dots, x_q) = (x_{q+1}, \dots, x_{2q})$ for $q = 1, 2, 4, 8, \dots$ recursively.

On the whole we see that pointer time $O(m^\beta N)$ is sufficient for the approximate computation of the characteristic polynomial of A such that the eigenvalue perturbations will be bounded by $O(2^{-N/m})$. Therefore, final application of (3.7) yields the following result (cf. [38]).

THEOREM. *Approximate determination of the eigenvalues of any complex $m \times m$ matrix (of norm ≤ 1) within error 2^{-n} is possible in pointer time*

$$O(m^\beta mn) + O(m^3 \log m + m^2 mn) \quad (\text{for any } \beta > \omega). \quad (4.8)$$

By comparing the two terms in this time bound we see (even with β close to two) that finding the eigenvalues from the characteristic polynomial is always faster than computing its coefficients, although the latter part does not cost much more than matrix multiplication, as far as the exponents are concerned.

The factor mn stands for the order of the precision N required under worst case assumptions. In practice, however, one will clearly try to get along with lower precision first, hoping for better a posteriori estimates which then may suffice to justify the low precision.

REFERENCES

1. A. Alder and V. Strassen, *On the algorithmic complexity of associative algebras*, Theoret. Comput. Sci. **15** (1981), 201–211.
2. W. Baur and V. Strassen, *The complexity of partial derivatives*, Theoret. Comput. Sci. **22** (1983), 317–330.
3. D. Bini, M. Capovani, G. Lotti, and F. Romani, $O(n^{2.7799})$ complexity for matrix multiplication, Inform. Process. Lett. **8** (1979), 234–235.
4. A. Borodin and I. Munro, *Computational complexity of algebraic and numeric problems*, American Elsevier, New York, 1975.
5. A. L. Chistov and D. Yu. Grigoryev, *Polynomial-time factoring of the multivariable polynomials over a global field*, LOMI preprint E-5-82, Leningrad, 1982.

6. ———, *Subexponential-time solving systems of algebraic equations*. I, II, LOMI Preprints E-9-83, E-10-83, Leningrad, 1983.
7. T. J. Chou and G. E. Collins, *Algorithms for the solution of systems of linear diophantine equations*, SIAM J. Comput. **11** (1982), 687–708.
8. G. E. Collins and R. Loos, *Real zeros of polynomials*, Computing, Suppl. **4** (1982), 83–94.
9. S. Cook, *A taxonomy of problems with fast parallel algorithms*, Inform. and Control **64** (1985), 2–22.
10. D. Coppersmith and S. Winograd, *On the asymptotic complexity of matrix multiplication*, SIAM J. Comput. **11** (1982), 472–492.
11. B. Dejon and P. Henriци (Editors), *Constructive aspects of the fundamental theorem of algebra* (Proc. Sympos., Zürich-Rüschlikon, 1967), Wiley, London, 1969.
12. L. M. Delves and J. N. Lyness, *A numerical method for locating zeros of an analytic function*, Math. Comp. **21** (1967), 543–560.
13. J. von zur Gathen, *Hensel and Newton methods in valuation rings*, Math. Comp. **42** (1984), 637–661.
14. P. Henriци, *Applied and computational complex analysis*, vol. 1, Wiley-Interscience, New York, 1974.
15. E. Kaltofen, *Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization*, SIAM J. Comput. **14** (1985), 469–489.
16. R. Kannan, *Solving systems of linear equations over polynomials*, Theor. Comput. Sci. **39** (1985), 69–88.
17. W. Keller-Gehrig, *Fast algorithms for the characteristic polynomial*, Theor. Comput. Sci. **36** (1985), 309–317.
18. D. E. Knuth, *The art of computer programming*, Vol. 2, Seminumerical Algorithms, 2nd ed., Addison-Wesley, Reading, Mass., 1981.
19. H. T. Kung, *On computing reciprocals of power series*, Numer. Math. **22** (1974), 341–348.
20. H. T. Kung and J. F. Traub, *All algebraic functions can be computed fast*, J. Assoc. Comput. Mach. **25** (1978), 245–260.
21. S. Landau, *Factoring polynomials over algebraic number fields*, SIAM J. Comput. **14** (1985), 184–195.
22. S. Landau and G. L. Miller, *Solvability by radicals is in polynomial time*, Proc. 15th Annual ACM Sympos. on Theory of Comp., Ass. Comp. Mach., New York, 1983, pp. 140–151.
23. D. Lazard, *Résolution des systèmes d'équations algébriques*, Theor. Comput. Sci. **15** (1981), 77–110.
24. A. K. Lenstra, *Factoring multivariate integral polynomials*, Theor. Comput. Sci. **34** (1984), 207–213.
25. ———, *Polynomial-time algorithms for the factorization of polynomials*, Ph.D. Thesis, Univ. of Amsterdam, 1984.
26. A. K. Lenstra, H. W. Lenstra, and L. Lovász, *Factoring polynomials with rational coefficients*, Math. Ann. **261** (1982), 515–534.
27. T. Lickteig, *The computational complexity of division in quadratic extension fields*, SIAM J. Comput. (to appear).
28. ———, *Gaussian elimination is optimal for solving linear equations in dimension two*, Inform. Process Lett. **22** (1986), 277–279.
29. A. Ostrowski, *Solution of equations and systems of equations*, 2nd ed., Pure and Applied Mathematics, vol. 9, Academic Press, New York, 1966.
30. V. Ya. Pan, *How can we speed up matrix multiplication?*, SIAM Rev. **26** (1984), 393–415.
31. A. Schönhage, *Schnelle Berechnung von Kettenbruchentwicklungen*, Acta Inform. **1** (1971), 139–144.
32. ———, *Unitäre Transformationen grosser Matrizen*, Numer. Math. **20** (1973), 409–417.

33. —, *Fast Schmidt orthogonalization and unitary transformations of large matrices*, Complexity of Sequential and Parallel Numerical Algorithms (J. F. Traub, ed.), Academic Press, New York, 1973, pp. 283–291.
34. —, *Schnelle Multiplikation von Polynomen über Körpern der Charakteristik 2*, Acta Inform. **7** (1977), 395–398.
35. —, *Storage modification machines*, SIAM J. Comput. **9** (1980), 490–508.
36. —, *Partial and total matrix multiplication*, SIAM J. Comput. **10** (1981), 434–455.
37. —, *Asymptotically fast algorithms for the numerical multiplication and division of polynomials with complex coefficients*, Computer Algebra (Marseille, 1982), Lecture Notes in Comput. Sci., Vol. 144, Springer, Berlin-New York, 1982, pp. 3–15.
38. —, *The fundamental theorem of algebra in terms of computational complexity*, Technical Report, Univ. Tübingen, 1982, 74 pp.
39. —, *Factorization of univariate integer polynomials by diophantine approximation and an improved basis reduction algorithm*, Automata, Languages and Programming (ICALP, Antwerp, 1984), Lecture Notes in Comput. Sci., Vol. 172, Springer, Berlin-New York, 1984, pp. 436–447.
40. —, *Quasi-GCD computations*, J. of Complexity **1** (1985), 118–137.
41. A. Schönhage and V. Strassen, *Schnelle Multiplikation grosser Zahlen*, Computing **7** (1971), 281–292.
42. M. Sieveking, *An algorithm for division of power series*, Computing **10** (1972), 153–156.
43. S. Smale, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (N.S.) **4** (1981), 1–36.
44. —, *On the efficiency of algorithms of analysis*, Bull. Amer. Math. Soc. (N.S.) **13** (1985), 87–121.
45. E. Specker and V. Strassen, *Komplexität von Entscheidungsproblemen*, Lecture Notes in Comput. Sci., Vol. 43, Springer, Berlin-New York, 1976.
46. V. Strassen, *Gaussian elimination is not optimal*, Numer. Math. **13** (1969), 354–356.
47. —, *Vermeidung von Divisionen*, J. Reine Angew. Math. **264** (1973), 184–202.
48. —, *Die Berechnungskomplexität von elementarsymmetrischen Funktionen und von Interpolationskoeffizienten*, Numer. Math. **20** (1973), 238–251.
49. —, *The computational complexity of continued fractions*, SIAM J. Comput. **12** (1983), 1–27.
50. —, *Algebraische Berechnungskomplexität*, Perspectives in Mathematics, Anniversary of Oberwolfach 1984 (W. Jäger, J. Moser, and R. Remmert, eds.), Birkhäuser, Basel, 1984, pp. 509–550.
51. —, *Relative bilinear complexity and matrix multiplication*, Manuscript, Universität Zürich, 1986.
52. P. Turán, *Power sum method and an approximative solution of algebraic equations*, Math. Comp. **29** (1975), 311–318.
53. H. Weyl, *Randbemerkungen zu Hauptproblemen der Mathematik. II, Fundamentalsatz der Algebra und Grundlagen der Mathematik*, Math. Z. **20** (1924), 131–150.

UNIVERSITÄT TÜBINGEN, D-7400 TÜBINGEN, FEDERAL REPUBLIC OF GERMANY

Taxonomy of Universal and Other Classes

SAHARON SHELAH

1. The problem. I was attracted to mathematics by its generality, its ability to give information where apparently total chaos prevails, rather than by its ability to give much concrete and exact information where we a priori know a great deal. So, not surprisingly, the following represents a theme which has been central in my mathematical interests since starting my thesis. (We give “universal classes” as an example, as the definition is “logic-free” (see Definition).)

1.1. *The first problem: The taxonomy = classification problem for universal classes.* Find the main dividing lines in the family of universal classes; each line is significant in the sense that all classes in one side “enjoy” common properties witnessing “simplicity,” “analyzability,” and those of the other side have common properties witnessing complications, unanalyzability; we define

1.2. *Universal classes.* (1) Examples are the class of groups, the class of rings, any variety, and the class of locally finite groups.

(2) Generally, let τ denote a vocabulary = set of function symbols and predicates (= relation symbols) each with an assigned arity, $(n(F), n(R))$; a τ -structure M is a nonempty set $|M|$ (its universe), and interpretations of any function symbol $F \in \tau$ and relation symbol $R \in \tau$ is an $n(F)$ -place function from $|M|$ to $|M|$, and $n(R)$ -place relation on $|M|$, respectively.

(3) A universal class K is a class of τ -structures for some $\tau = \tau(K)$ such that $M \in K$ iff every finitely generated substructure of M belongs to K .

I also love, in mathematics, that there is no argument (at least usually) about whether or not one solves a problem. It suffices to find the correct solution; being untalented in convincing people is no serious hindrance. Hence I like problems that are precise, preferably with a yes/no answer, and I believe that usually the way to treat more elusive problems is by choosing the right test question. So we shall specify our problem below.

1.3. *The second problem: The structure/nonstructure problem.* (1) Describe for some (e.g., universal classes) K a structure theory (see below) and prove for the other classes (in the family) nonstructure theorems; that is, demonstrate

This research was partially supported by the United States–Israel Binational Science Foundation and the National Science Foundation.

the impossibility of a structure theory, construction of many and/or complicated structures in K .

(2) A *structure theory* for K is a theory that gives, for each $M \in K$, a complete set of invariants; i.e., each invariant should depend only on the isomorphism type of M , and if M_1, M_2 have the same invariants, then they are isomorphic.

Of course, we do not want the invariants to be too complicated (e.g., the isomorphism type of M) (and, preferably, derivation of the invariant from the structure and vice versa are explicit constructions). We shall not deal with this, but it would not change the theory much. See [Sh86] for more.

What objects should we use as invariants for structures of cardinality λ ? In the prototype of structure theorems, the celebrated Steinitz theorem, for algebraically closed fields of a fixed characteristic, a cardinality (= transcendence dimension) is used. Certainly if K_i has a structure theory for $i \in I$, then so does $\sum_{i \in I} K_i \stackrel{\text{df}}{=} \{\sum_{i \in I} M_i : M_i \in K_i \text{ for } i \in I\}$ (with any reasonable definition of $\sum_i M_i$, such that $\sum_{i \in I} K_i$ is a class of the right kind). Also, if K has a structure theory so does $\sum K \stackrel{\text{df}}{=} \{\sum_{j \in J} M_j : \text{for some set } J, \text{ with } M_j \in K\}$. So we have to admit λ -values of kind (α, χ) as invariant for structures of cardinality λ , where

1.4. *Definition.* (1) A λ -value of kind (α, χ) (λ, χ cardinals, α an ordinal) is defined by induction on α :

$\alpha = 0$. A λ -value of kind (α, χ) is a cardinal $\leq \lambda$.

$\alpha = \beta + 1$. A λ -value of kind (α, χ) is a λ -value of kind (β, χ) or sequence of length χ , each entry a function from $\{x : x \text{ a } \lambda\text{-value of kind } (\beta, \chi)\}$ into the set of cardinals $\leq \lambda$.

α limit. A λ -value of kind (α, χ) is a λ -value of kind (β, χ) for some $\beta < \alpha$.

(2) We say an invariant of kind $(\beta, 1)$ is of depth β .

(3) An invariant of kind (α, χ) , for K , is a function which gives, for each $M \in K$ of cardinality λ , a λ -value of kind (χ, α) , and which depends on M only up to isomorphism.

Now if we do not bound α , even for $\chi = 1$, any structure of cardinality λ can be coded up to isomorphism; also note that for any χ, α there is a β such that every λ -value of kind (χ, α) can be coded naturally by a λ -value of depth β (i.e., kind $(\beta, 1)$). So we are led to

1.5. *First thesis:* A class K has a structure theory iff for some β there are invariants of depth β for K which determine each $M \in K$ up to isomorphism.

I think that the part of the thesis that says that such invariants give a structure theory is very strong. First of all

1.6. *Claim.* If a class K has a structure theory according to Thesis 1.5, by an invariant of kind (α, χ) , then for every cardinal \aleph_γ , $I(\aleph_\gamma, K) \leq \beth_\alpha(|\gamma| + \chi)$ (see below), so if the Generalized Continuum Hypothesis (= G.C.H.) holds, then $\beth(\aleph_\gamma, K) \leq (\chi + |\gamma|)^{+\beta}$ where

1.7. *Definition.* (1) $|\gamma|$ is the cardinality of the ordinal γ .

(2) $I(\aleph_\gamma, K)$ is the number of structures from K of cardinality \aleph_γ , up to isomorphism.

(3) The G.C.H. says that $2^{\aleph_\gamma} = \aleph_{\gamma+1}$ for every γ .

(4) $(\aleph_\alpha)^{+\beta} = \aleph_{\alpha+\beta}$, $\beth_\alpha(\lambda)$ is defined by induction on α as $\aleph_0 + \sum_{\beta < \alpha} 2^{\beth_\beta(\lambda)}$.

(Note that $\beth_\alpha(\chi + |\gamma|)$ may be $> 2^{\aleph_\gamma}$ even for $\alpha = 2$, $\chi = 1$.)

Because knowing the number of members of K in one cardinality is a natural and important problem, $I(\lambda, K)$ is an important function; if its values are small this signifies that the class K is simple. For me, the really important thing is Thesis 1.8 below (and not the detailed computation as $I(\lambda, K)$).

1.8. *Second thesis: The (first) main dividing line = the main gap.* For a “nice” family of classes (like the family of universal classes), the dividing line “is there α such that for every γ , $I(\aleph_\gamma, K) \leq \beth_\alpha(|\gamma|)$ ” is a good one, i.e.,

(a) It coincides with “having a structure theory according to 1.5”;

(b) Every class in the “complicated side” has strong evidence for nonstructure: a jump in the lower bound for $I(\lambda, K)$ —i.e., among the possible functions $I(\lambda, K)$ there is a large gap:

(1) $I(\lambda, K) = 2^\lambda$ for large enough λ which is the maximum value when $\lambda \geq |\tau(K)|$, or at least

(2) $I(\lambda, K) > \lambda$ for large enough class of cardinals λ . We much prefer that the nonstructure proof be carried out in ZFC alone, but note that in order to show that we cannot prove a structure theorem, the consistency of nonstructure is enough.

(c) We believe that if we succeed in solving (b), we will have developed extended taxonomy having many tools to deal with, and we will know much on each side of those dividing lines; i.e., we suggest 1.8(b) as the test question.

Note that we do not claim that only dividing lines are interesting: the class of rings is a very interesting subclass of the class of structures with two-place functions, although its complement is not interesting at all. On the other hand, dividing lines, in addition to having intrinsic interest, help in proving theorems by cases.

Implicit is

1.9. *Thesis.* (d) In understanding a class you should look at a large enough cardinality in order to iron out singularities.

Note that, e.g., theories having unique countable model can have many complicated uncountable ones. Note that a priori the answer to the following question is not clear.

1.10. *Question.* Is there a reasonably general family of classes for which Thesis 1.8 can be confirmed?

We shall return to this question later.

2. Background. Why do we speak on universal classes? Now in model theory the primary family of classes is the family $\{\text{Mod}(T) : T \text{ a countable first-order theory}\}$, where

2.1. *Definition.* $\text{Mod}(T)$ is the class of models of T (and we write T instead of $\text{Mod}(T)$).

Of course, more complicated classes—uncountable theories, theories in infinitary logics or ones with generalized quantifiers on the one hand, and universal theories and even varieties (= equational theories)—are also interesting. I think this approach is right, but here there is no need to justify it.

Let us return to what is for me prehistory (you can ignore notions unknown to you). Generally, around 1960, research in mathematical logic became deeper and more complicated mathematically. At that time, the aim of much of the research in model theory was advancement toward the solution of the Los Conjecture, which was as follows.

LOS CONJECTURE. *If T is (first-order) countable, then if T is categorical in λ for some $\lambda > \aleph_0$ (i.e., $I(\lambda, T) = 1$), then this holds for every $\lambda > \aleph_0$.*

Certainly, in wanting to know something about $I(\lambda, T)$, this was a very good problem to start with: it was foolish to consider far-reaching conjectures like 1.8(b), considering the knowledge available. Those investigations culminated in the positive solution of the Los Conjecture by Morley in [Mo], which used many of the tools developed previously—in particular, the works of Vaught and Ehrenfeucht-Mosłowski. Morley's theorem is considered by many (including myself) to be one of the main achievements of mathematical logic during the sixties.

For quite some time, little happened. This certainly has its reasons—among them, that Morley and Keisler, at least, thought that the (model) theory of first-order theories was finished, or essentially finished (they told me so in 1969). However, since then the field has increased in popularity in model theory, as witnessed by the research book [Sh78], and numerous articles, as well as the (largely) expository books Pillay [Pi], Lascar [La], Poizat [Po], and Baldwin [Ba], and lately some conferences dedicated to it. On early history see [Sh74]; this article overlaps with [Sh85] (which speaks on countable first-order T) and Baldwin, introduction to [Ba1] (which deals generally with the theory).

The papers of most researchers in the field, however, reflect a very different outlook—a “fine structure” one—wanting to know much (or everything) about what we already have considerable knowledge about or investigating families of classes which have some structure theory and/or tendency to be (relatively) more concrete. Illustrious examples are the Baldwin-Lachlan theorem (on first-order countable T , categorical in \aleph_1 but not \aleph_0) and the works of Zilber and Cherlin, Harrington and Lachlan (on T categorical in \aleph_0 , \aleph_1 , or totally transcendental T categorical on \aleph_0). I hope Lachlan and Peretyatkin, in this volume, will do justice to some of this.

Note that there is no real conflict: solutions of Problem 1.1 give, and are intended to give, instances for fine structure investigation with considerable tools to start with. Here I shall continue to present my personal outlook. When I started in 1967, I was interested in Problem 1.1, introducing the stability and superstability as dividing lines, and in [Sh71, p. 283, (13)] (also in [Sh74]) I conjectured what the functions $I(\lambda, T)$ for λ large enough, T countable first-order, should be like; in particular, the “main gap” of 1.7(b) was of interest.

The first class for which this was confirmed (with a minor correction) was for the family of $\{M: M \text{ an } \aleph_\varepsilon\text{-saturated model of } T\}$: T first-order (announced in [Sh74], see [Sh83; Sh83a; Sh77, Chapter X, Example 3.3]). but I felt this was cheating, as I invented \aleph_ε -saturation. The second one was for universal (first-order) theories T (see [Sh86]); i.e., the set of models of T , T consisting of formulas of the form $\forall x_1, \dots, x_n \bigvee_i \Lambda_j \phi_{i,j}$, $\phi_{i,j}$ atomic or negation of atomic. We will give more details on this result.

2.2. THEOREM. *For every universal (first-order) T , exactly one of the following occurs:*

(A) *For every $\lambda > |T|$, $I(\lambda, T) = 2^\lambda$, and there are other signs of complication (see [Sh85]).*

(B) *Still $I(\lambda, T) = 2^\lambda$ for every $\lambda \geq |T| + \aleph_1$, but*

(*) *for every model M of T of cardinality λ , there is $\langle M_\eta : \eta \in I \rangle$ such that:*

(i) *I is a set of finite sequences of ordinals $< \lambda$, nonempty, closed under initial segments.*

(ii) *M_η is a submodel of M of cardinality $\leq |T|$, and if ν is an initial segment of η then M_ν is a submodel of M_λ .*

(iii) *M is freely generated by $\bigcup_{\eta \in I} M_\eta$, which means:*

(α) *the closure of $\bigcup_{\eta \in I} |M_\eta|$ (union of universes) under the functions of M is $|M|$;*

(β) *if $\eta \in I$ has length $n+1$, $\nu = \eta \restriction n$, then for every finite sequence \bar{c} from $M_\eta : \oplus$ for every finite set Φ of quantifier-free formulas with parameters from $\bigcup\{|M_\rho| : \rho \in I, \eta \text{ not an initial segment of } \rho\}$ which \bar{c} satisfies, there is a finite sequence \bar{c}' from M_ν satisfying all formulas in Φ .*

(C) *The condition (*) from (B) holds, moreover, for some ordinal $\text{Dp}(T)$, which is countable if T is countable, and of cardinality $\leq 2^{|T|}$ generally; for every M there is $\langle M_\eta : \eta \in I \rangle$ as in (*) with I of depth $\langle \text{Dp}(T) : \text{i.e., there is a function } d: I \rightarrow \alpha \text{ such that if } \nu \text{ is a proper initial segment of } \eta \text{ then } d(\nu) > d(\eta). \rangle$*

So for α large enough, $I(\aleph_\alpha, T) \leq I_{\text{Dp}(T)}(|\alpha|)$ (in fact, we get very detailed information on $I(\lambda, T)$).

Note that the first dividing line is between (A)+(B) and (C). But a second dividing line between (A) and (B)+(C) is worthwhile; see [Sh85] on this, but we shall not deal with it here.

This looks to be a reasonable answer to Question 1.10—a general family where we have a solution; but, having a model theoretic background, I was not satisfied until the solution for countable first-order T (see [Sh85, Sh87]). Though I thought 1.8 was the point, I felt it was my duty not to avoid the relevant problem which was the legacy of the previous generation—the Morley Conjecture. Having thus answered Question 1.10 fully, we have

2.3. Question. *Is the theory a theory on first-order classes, or is there really a collection of such theories which may even need a general framework?*

There were some advances, some changing of the family of classes, some changing of the questions. In Baldwin-Shelah [BaSh8] (see also Shelah [Sh86a, Sh85a]), the complexity of class $K = \text{Mod}(T)$, T first-order, in the monadic logic (finitary or infinitary) was investigated (relying on [Sh78]).

Quite complete classification (in the relevant sense) was obtained. The results explain why the quite large body of works on monadic logic (e.g., works of Buchi, Rabin, Shelah, and Gurevich (see [Gu])) concentrate on linear orders and trees, and discover some neglected cases; also, more general quantifiers were dealt with. This also indicates that the classification in [Sh78] is relevant to a reasonably wide spectrum of problems not thought of in the first place. On classifying all quantifiers see [Sh86x].

We may work with classes K of two sorted structures and ask how much a $M \in K$ can be described up to isomorphism over the first sort. The prototype of such problems is the structure of a vector space V over a field F , letting both vary; so one cardinal invariant, the dimension, suffices.

No nonstructure will mean that for each (or at least arbitrarily large) cardinal λ , there is a structure M_0 so that for many and complicated $N \in K$, N restricted to the first sort is M_0 . It seems that for $K = \text{Mod}(T)$, T countable first-order, there is a complete answer *provided that* we accept indepenent results for the nonstructure side. We say “it seems,” as some parts are not worked out, others need considerable expansion; see [PiSh8], [Sh86], and the notes [Sh85b]. However, that is enough to show that there is such a theory.

The situation is similar for universal classes [Sh87a].

Grossberg and Hart [GH] have been proving the main gap, etc., for the family of excellent classes. Excellent classes were introduced in [Sh83c] (in the following context: if $\psi \in L_{\omega_1, \omega}$ has an uncountable model and, for no $n > 0$, has many nonisomorphic models (essentially 2^{\aleph_n}), then $\text{Mod}(\psi)$ is the union of few excellent classes).

3. Outside interactions. Of course, the theory answered almost all relevant problems of the model theorist of the sixties and some which do not a priori look connected (like investigating Keisler order on first-order theories).

Note that the conclusion applies also to universal algebra. In particular, Theorem 2.2 gives the possible function $I(\lambda, K)$ for K a variety, except that we do not know if all values of the parameters really appear (mainly whether the depth of the theory can be infinite; also there are some problems for small cardinals). It seemed that though they have interest in the problem, universal algebraists have not learned the material we mention.

We have been interested in this theory for its own sake; and applications were sought in order to convince the “heathen.” However, we sincerely believe that it should help in investigating specific classes.

Note, however, that there is an asymmetry between the structure and non-structure side. You can deal with a structure theory for a specific class without

having any formal definition of what a structure theory is. But we need one (or some alternatives) for showing that no structure theory exists.

Note that having a theory as we desire makes it natural and profitable to investigate for specific classes where they are on this classification. This line of research starts with Macintyre [Mc], dealing with first-order theories of fields (via [Mo]). For field and division rings this was continued in Cherlin [Ch] and Cherlin and Shelah [ChSh], thus giving the following

Conclusion. If T is a first-order theory, $\text{Mod}(T)$ a class of infinite fields (or division rings), not all algebraically closed fields, then T is not superstable; hence $\text{Mod}(T)$ has many and complicated models and modules.

There is extensive literature on first-order theories of rings, modules and groups. A very successful case is the theory T_{dcl}^p of differentially closed fields. Robinson, relying on Seidenberg [Se], proved that T_{dcl}^0 is first-order. Blum [B] gave concrete axioms for it, and proved it totally transcendental. She deduced (relying on Morley's work) that over every differential field F of characteristic 0, there is a prime differentially closed field over F (extending F) (i.e., one embeddable in every differentially closed field extending F). We deduce, by a theorem in classification theory (see [Sh78, Chapter IV, §4]), the uniqueness of the prime differentially closed field over F . Wood [W], relying again on algebraic work, proved T_{dcl}^p is first-order, but not totally transcendental. Independently, Shelah [Sh] and Wood [W1] proved the existence of prime differentially closed fields over any differential field. Shelah [Sh73] proved that the theory is stable; so by [Sh78, Chapter IV, §5], the prime differentially closed field above is unique. But by [Sh73], T_{dcl}^p ($p > 0$) is not superstable, hence there is no structure theory for $\text{Mod}(T_{\text{dcl}}^p)$.

The existing theory on $\text{Mod}(\psi)$, ψ a sentence in infinitary logic ($L_{\omega_1, \omega}$), is used in Mekler and Shelah [MSh] to prove (when $V = L$) that for every variety either $L_{\infty, \omega}$ -freeness implies freeness or there are λ -free not free ones for every λ (continuing work of Eklof and Mekler [EM]).

In Grossberg and Shelah [GSh] a problem of Fuchs and Salce (see [FS]) on the possibility of a structure theory of torsion divisible modules over a uniserial ring is answered (using a general theorem proved there). This theorem can also be used to deduce directly an older result from [Sh74a] (solving a problem of Fuchs [F]) that there are many complicated separable reduced abelian p -groups in every $\lambda > \aleph_0$.

Quite naturally, in many cases the theorems have not applied directly; rather, the proofs or the method apply. We have tried to adapt the theorems to general use in [Sh83b], the applications there being constructions of Boolean algebras which are complicated in various ways (e.g., have no automorphisms or one-to-one endomorphisms, are complete and/or satisfy the CCC).

Another attempt to adapt the theorems for applications is [Sh85c], which has been of use in several instances in representing rings as endomorphism groups of abelian groups, in works of Corner, Gobel, and the author.

We generally believe that the method should be useful in constructing structures in specific classes which are "complicated," e.g., have no "nontrivial" automorphism or endomorphism or are indecomposable, etc.

REFERENCES

- [B] L. Blum, *Generalized algebraic structures, a model theoretic approach*, P.D. Thesis, M.I.T., Cambridge, Mass., 1968.
- [Ba] J. Baldwin, Springer Verlag.
- [Ba1] J. Baldwin (ed.), Introduction, USA-Israel, Proc. Conference on Classification Theory (Chicago, December 1985), Lecture Notes in Math., Springer-Verlag, Berlin and New York (to appear).
- [Ba2] J. Baldwin, *Definable second order quantifiers*, Model-Theoretic Logics (J. Barwise and S. Feferman, eds), Springer-Verlag, Berlin and New York, 1985, pp. 445–478.
- [BaSh] J. Baldwin and S. Shelah, *Classification of theories by second order quantifiers*, Proc. 1980/Jerusalem Model Theory Year, Notre Dame J. Formal Logic **26** (1985), 229–303.
- [Ch] G. Cherlin, *Superstable division rings*, Logic Colloquium 77 (Proc. Conf., Wrocław, 1977), North-Holland, Amsterdam, 1978, pp. 99–112.
- [ChSh] G. Cherlin and S. Shelah, *Superstable fields and groups*, Ann. of Math. Logic **18** (1980), 227–280.
- [EM] P. Eklof and A. Mekler, *Categoricity results for $L_{\infty\kappa}$ -free algebras*, Ann. Pure Appl. Logic (to appear).
- [F] L. Fuchs, *Infinite Abelian groups*, Academic Press, 1970, 1973.
- [FS] L. Fuchs and L. Salce, *Modules over valuation domains*, Lecture Notes in Pure Appl. Math., vol. 97, Marcel Dekker, New York, 1985.
- [Gu] Y. Gurevich, *Monadic second order theories*, Model-Theoretic Logics (J. Barwise and S. Feferman, eds.), Springer-Verlag, Berlin and New York, 1985, pp. 479–506.
- [GSh] R. Grossberg and S. Shelah, *A nonstructure theorem for an infinitary theory which has the unsuperstability property*, Illinois J. Math. **30** (1986), 364–390.
- [K] E. R. Kolchin, *Differential algebra and algebraic groups*, Academic Press, New York, 1973.
- [La] D. Lascar, *Introduction to stability*.
- [Mc] A. Macintyre, *On ω_1 -categorical theories of fields*, Fund. Math. **71** (1971), 1–25.
- [Mo] M. D. Morley, *Categoricity in power*, Trans. Amer. Math. Soc. **114** (1965), 514–538.
- [MSh] A. Mekler and S. Shelah, *For which varieties $L_{\infty, \omega}$ -freeness implies freeness and excellent classes*, in preparation.
- [Pi] A. Pillay, *An introduction to stability theory*, Clarendon Press, Oxford, 1983.
- [PiSh] A. Pillay and S. Shelah, *Classification over a predicate*. I, Notre Dame J. Formal Logic **26** (1985), 361–376.
- [Po] B. Poizat, *Cours de théorie des modèles*, nur al-mantiq wal-ma'rifah, 1985.
- [Ro] A. Robinson, *On the concept of a differentially closed field*, Bulletin Research Council of Israel, Section F (later: Israel J. Math) **8F** (1959), 113–128.
- [Se] A. Seidenberg, *An elimination theory for differential fields*, Univ. Calif. Publ. Math. (N.S.) **3** (1956), 31–65.
- [Sh71] S. Shelah, *Stability, the f.c.p. and superstability, model theoretic properties of formulas in first order theory*, Ann. of Math. Logic **3** (1971), 271–362.
- [Sh73] —, *Differentially closed fields*, Israel J. Math. **16** (1973), 314–328.
- [Sh74] —, *Categoricity of uncountable theories*, Proc. Sympos. Pure Math., vol. 25, Amer. Math. Soc., Providence, R.I., 1974, pp. 187–204.
- [Sh74a] —, *Infinite abelian groups, Whitehead problem and some constructions*, Israel J. Math. **18** (1974), 243–256.
- [Sh78] —, *Classification theory and the number of nonisomorphic models*, North-Holland, Amsterdam, 1978.
- [Sh83] —, *The spectrum problem*. I, \aleph_ε -saturated models the main gap, Israel J. Math. **43** (1982), 324–356.

- [Sh83a] —, *The spectrum problem. II, Totally transcendental theories and the infinite depth case*, Israel J. Math. **43** (1982), 357–364.
- [Sh83b] —, *Construction of many complicated uncountable structures and Boolean algebras*, Israel J. Math. **45** (1983), 100–146.
- [Sh83c] —, *Classification theory for non-elementary classes. I, The number of uncountable models, models of $\psi \in L_{\omega_1, \omega}$* , Israel J. Math. **46** (1983), 2–12–273.
- [Sh85] —, *Classification of first order theories which have a structure theory*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), 227–232.
- [Sh85a] —, *Monadic Logic: Lowenheim numbers*, Ann. Pure Appl. Logic **28** (1985), 203–216.
- [Sh85b] —, *Classification over a predicate*, Notes from Lectures in Simon Fraser University, Summer 1985.
- [Sh85c] —, *A combinatorial principle and endomorphism rings of abelian groups. II*, Proc. of the Conference on Abelian Groups Indine 4/1984, CISM courses and Lecture ∞ , No. 287, International Center for Mechanical Sciences, *Abelian Groups and Modules*, R. Gobel, C. Metelli, A. Orsatti, and L. Salce, eds., 1985, pp. 37–86.
- [Sh86] —, *Spectrum problem. III, Universal theories*, Israel J. Math. **55** (1986), 229–252.
- [Sh86a] —, *Monadic logic: Hanf numbers*, Around Classification Theory, Lecture Notes in Math., vol. 1182, Springer-Verlag, Berlin and New York, 1986, pp. 203–223.
- [Sh86b] —, *Classification over a predicate. II*, Around Classification Theory, Lecture Notes in Math., vol. 1182, Springer-Verlag, Berlin and New York, 1986, pp. 47–90.
- [Sh86c] —, *Classifying generalized quantifiers*, Around Classification Theory, Lecture Notes in Math., vol. 1182, Springer-Verlag, Berlin and New York, 1986, pp. 1–46.
- [Sh87] —, *Classification theory; completed for countable theories*, North-Holland, Amsterdam (to appear).
- [Sh87a] —, *Universal classes*, Proc. of the USA-Israel Sympos. on Classification Theory, Chicago 12/85, Springer-Verlag.
- [W] C. Wood, *The model theory of differential fields of characteristic $p \neq 0$* , Proc. Amer. Math. Soc. **40** (1973), 577–584.
- [W1] —, *Prime model extensions for differential fields of characteristic $p \neq 0$* , J. Symbolic Logic **39** (1974), 469–477.

THE HEBREW UNIVERSITY, JERUSALEM, ISRAEL

RUTGERS UNIVERSITY, NEW BRUNSWICK, NEW JERSEY 08903, USA

Случайные процессы в бесконечномерных пространствах

А. В. СКОРОХОД

В последние годы в теории случайных процессов все больше внимания уделяется изучению случайных процессов в бесконечномерных пространствах. Интерес к таким процессам вызван не только желанием распространить известные методы исследования конечномерных процессов, но и тем, что бесконечномерные процессы возникают и при решении ряда задач, естественных для конечномерных процессов. Те особенности и трудности, которые возникают при рассмотрении бесконечномерных процессов, связаны в первую очередь с геометрической структурой пространства. Поэтому такие трудности исчезают, если удастся найти подходящее отображение исходного пространства в некоторое другое. Заметим, что упомянутые трудности появляются при рассмотрении вопросов конструктивного плана (изучение свойств выборочных функций, исследование сходимости случайных процессов, построение стохастических интегралов, определение и исследование решений стохастических дифференциальных уравнений). Общие вопросы теории меры безразличны к топологии.

В работе мы отвлекаемся от топологии фазового пространства и исследуем случайный процесс в измеримом пространстве со счетно-порожденной σ -алгеброй. Такие пространства составляют достаточно обширный класс, только такие пространства могут использоваться в приложениях. Выясняется структура измеримого процесса. Определяются некоторые инварианты процесса, которые при весьма широких ограничениях определяют процесс с точностью до неупреждающих обратимых преобразований. Среди этих инвариантов есть ранг процесса, который может быть равен или $+\infty$, или натуральному числу. Ранг и есть существенная размерность процесса. Случайной заменой времени и неупреждающим преобразованием процесс можно свести к процессу в R^∞ , удовлетворяющему линейному стохастическому

дифференциальному уравнению типа Ито. Ранг процесса — это число входящих в уравнение независимых винеровских процессов.

В работе без специальных ссылок используются результаты Мейера, Кунита, Ватанабе, Делашери по теории мартингалов и общей теории случайных процессов. Эти результаты широко известны и имеются в книгах этих авторов и во многих других книгах по теории случайных процессов.

1. Теорема Колмогорова. Случайный процесс определяется такими тремя объектами: (1) вероятностным пространством (Ω, \mathcal{F}, P) , (2) фазовым пространством (X, \mathcal{B}) , которое является измеримым пространством, (3) отображением $x(t, \omega): R_+ \times \Omega \rightarrow X$, обладающим тем свойством, что для всех $t \in R_+$ отображение $x(t, \cdot): R_+ \rightarrow X$ измеримо относительно σ -алгебр \mathcal{F} и \mathcal{B} . Функция $x(t, \omega)$ и есть случайный процесс в фазовом пространстве X , определенный на неотрицательной полупрямой R_+ . Вероятностное пространство играет существенную роль лишь в том случае, когда процесс определяется с помощью некоторой конструкции. При общем рассмотрении случайных процессов основной характеристикой процесса являются конечномерные распределения. Это семейство функций

$$\begin{aligned} \mu_{t_1, \dots, t_n}(B_1, \dots, B_n) &= P\{x(t_1, \omega) \in B_1, \dots, x(t_n, \omega) \in B_n\}, \\ n &= 1, 2, \dots, \quad t_k \in R_+, \quad B_k \in \mathcal{B}. \end{aligned} \quad (1)$$

Эти функции удовлетворяют следующим условиям согласования:

(1)

$$\begin{aligned} \mu_{t_{i_1}, \dots, t_{i_n}}(B_{i_1}, \dots, B_{i_n}) &= \mu_{t_1, \dots, t_n}(B_1, \dots, B_n), \\ \{i_1, \dots, i_n\} &= \{1, 2, \dots, n\}, \quad n = 1, 2, \dots, \quad t_k \in R_+, \quad B_k \in \mathcal{B}, \end{aligned}$$

(2)

$$\mu_{t_1, \dots, t_{n-1}, t_n}(B_1, \dots, B_{n-1}, X) = \mu_{t_1, \dots, t_{n-1}}(B_1, \dots, B_{n-1}),$$

(3)

$$\mu_{t_1, \dots, t_n}(B, B_2, \dots, B_n) \text{ есть мера по } B, \quad \mu_t(X) = 1.$$

Теорема Колмогорова утверждает, что для всякого набора согласованных конечномерных распределений существует случайный процесс $x(t, \omega)$, определенный на некотором вероятностном пространстве $\{\Omega, \mathcal{F}, P\}$, для которого выполнено (1).

В качестве Ω можно взять пространство X^{R_+} — пространство всех X -значных функций, определенных на R_+ ; при этом \mathcal{F} будет цилиндрической σ -алгеброй в X^{R_+} , а мера P однозначно определяется конечномерными распределениями равенством

$$P(\{x(\cdot): x(t_1) \in B_1, \dots, x(t_n) \in B_n\}) = \mu_{t_1, \dots, t_n}(B_1, \dots, B_n).$$

Сам процесс $x(t, \omega)$ задан соотношением

$$x(t, \omega(\cdot)) = \omega(t), \quad \omega(\cdot) \in X^{R_+}.$$

2. Измеримость. Пусть $x(t, \omega)$ и $\tilde{x}(t, \omega)$ два случайных процесса, заданных на одном и том же вероятностном пространстве. Они называются стохастически эквивалентными, если для всех $t \in R_+$

$$P\{x(t, \omega) = \tilde{x}(t, \omega)\} = 1.$$

Тогда еще $x(t, \omega)$ и $\tilde{x}(t, \omega)$ называются модификациями друг друга. При задании процесса конечномерными распределениями естественно не различать модификации. Дальнейшее изучение случайного процесса также не зависит от выбора модификации.

Будем называть процесс измеримым, если он имеет измеримую модификацию, т.е. такую модификацию $\tilde{x}(t, \omega)$, которая является измеримым отображением пространства $(R_+ \times \Omega, \mathcal{B}_{R_+} \otimes \mathcal{F})$ в (X, \mathcal{B}) , \mathcal{B}_{R_+} — борелевская σ -алгебра на R_+ .

Для дальнейшего нам потребуются некоторые пространства случайных величин. Обозначим через $R(\Omega)$ пространство всех числовых величин на вероятностном пространстве $\{\Omega, \mathcal{F}, P\}$. Если $\xi, \eta \in R(\Omega)$, то

$$\rho(\xi, \eta) = M(1 - e^{-|\xi - \eta|})$$

есть метрика в $R(\Omega)$, относительно которой $R(\Omega)$ — полное пространство.

Пусть H_t замыкание в этой метрике множества величин вида $g(x(s_1, \omega), \dots, x(s_n, \omega))$, где $n = 1, 2, \dots$, $s_k \in R_+$, $s_k \leq t$; $g(x_1, \dots, x_n)$ есть B^n -измеримая функция X^n в R .

H_t — величины, определяемые течением процесса до момента t включительно.

Через H обозначим замыкание $\bigcup_t H_t$.

ТЕОРЕМА 1. Пусть σ -алгебра \mathcal{B} счетно порождена. Процесс $x(t, \omega)$ измерим тогда и только тогда, когда: (1) H сепарабельно, (2) $\mu_{s,t}(B_1, B_2)$ — борелевская функция s при любых $t \in R_+$, $B_1, B_2 \in \mathcal{B}$.

В случае счетно порожденного (X, \mathcal{B}) это пространство можно измеримо взаимно однозначно отобразить в отрезок $[0, 1]$. Это тем не менее не дает основания утверждать, что по существу все процессы с такими фазовыми пространствами одномерны. В дальнейшем будем предполагать, что σ -алгебра \mathcal{B} счетно порождена, а процесс $x(t, \omega)$ является измеримым.

3. Определяющий поток σ -алгебр. Эволюция процесса во времени определяет изменение пространств H_t . Заметим, что эти пространства не меняются при взаимно однозначных неупреждающих преобразованиях случайного процесса, в частности взаимно однозначных отображениях фазовых пространств. Вместо семейства пространств H_t можно рассмотреть поток σ -алгебр (\mathcal{F}_t) , где \mathcal{F}_t — σ -алгебра событий, порожденная случайными величинами из H_t . Так как \mathcal{F}_t и H_t определяют друг друга, то будем изучать строение потока (\mathcal{F}_t) . Обозначим через M_2 пространство квадратически интегрируемых (\mathcal{F}_t) -согласованных мартингалов.

ТЕОРЕМА 2. (1) M_2 определяет поток (\mathcal{F}_t) , (2) существует счетное множество мартингалов $\{\eta_n(t), n = 1, 2, \dots\}$ в M_2 такое, что для каждого $t \in R_+$ значения мартингалов плотны в H_t .

Для дальнейшего нам потребуются следующие условия на поток (\mathcal{F}_t) :

- (1) \mathcal{F} полно относительно \mathcal{P} , F_t содержит все множества меры 0,
- (2) $\mathcal{F}_t = \bigcap \mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s$.

При выполнении этих условий мартингалы из M_2 имеют непрерывные справа модификации без разрывов второго рода. Можно считать, что M_2 состоит только из таких мартингалов. Будем обозначать через \mathcal{W} и \mathcal{P} σ -алгебры вполне измеримых и предсказуемых множеств в $R_+ \times \Omega$. Первая порождается непрерывными справа, а вторая — непрерывными согласованными процессами.

Для $\eta_1(t)$ и $\eta_2(t)$ из M_2 обозначим через $\langle \eta_1, \eta_2 \rangle_t$ их взаимную характеристику — такой \mathcal{P} -измеримый процесс, что $\eta_1(t)\eta_2(t) - \langle \eta_1, \eta_2 \rangle_t$ есть мартингал. Величина $\langle \eta, \eta \rangle_t = \langle \eta \rangle_t$ — характеристика мартингала $\eta(t)$. Два мартингала $\eta_1(t)$ и $\eta_2(t)$ из M_2 ортогональны ($\eta_1 \perp \eta_2$), если $\langle \eta_1, \eta_2 \rangle_t = 0$ для всех $t \in R_+$.

ТЕОРЕМА 3. $M_2 = M_2^0 \oplus M_2^1 \oplus M_2^2$ (ортогональная сумма в смысле введенного понятия ортогональности),

M_2^0 — пространство непрерывных мартингалов,

M_2^1 — пространство чисто разрывных компенсированных мартингалов с непрерывными характеристиками,

M_2^2 — пространство чисто разрывных мартингалов с чисто разрывными характеристиками.

Это разложение имеет аналогию в известном разложении Леви процесса с независимыми приращениями на чисто разрывную (M_2^2), стохастически непрерывную скачкообразную (M_2^1) и непрерывную составляющие. Как мы увидим дальше эта аналогия может быть продолжена более глубоко.

4. Существенное время. Ранг. Рассмотрим пространство $M_2^0 \oplus M_2^1$ всех мартингалов из M_2 с непрерывными характеристиками. Выбирая некоторую плотную последовательность мартингалов из $M_2^0 \oplus M_2^1$ и рассматривая их характеристики можно построить такой мартингал $\bar{\eta}(t) \in M_2^0 \oplus M_2^1$, что для всякого другого мартингала $\eta(t) \in M_2^0 \oplus M_2^1$ возрастающая функция $\langle \eta \rangle_t$ будет абсолютно непрерывна относительно возрастающей функции $\langle \bar{\eta} \rangle_t = \delta_t$. Всякую такую функцию δ_t (она представима как характеристика некоторого мартингала и относительно нее абсолютно непрерывны характеристики всех других мартингалов из $M_2^0 \oplus M_2^1$) будем называть существенным временем для потока (\mathcal{F}_t) (или исходного процесса). Если на некотором интервале δ_t не растет, то на этом интервале исходный процесс эволюционирует детерминированно, хотя эволюция зависит от того, что происходило до этого интервала

времени. Очевидно, что существенное время определяется с точностью до эквивалентности: если δ_t — существенное время и $\hat{\delta}_t$ непрерывный возрастающий согласованный процесс, для которого

$$P\{0 < d\hat{\delta}_t/d\delta_t < \infty\} = 1,$$

то $\hat{\delta}_t$ — также существенное время.

Рассмотрим теперь пространство M_2^0 . Пусть $\{\eta_n(t)\}$ плотно в M_2^0 . Ортогонализуем $\{\eta_n(t)\}$ следующим образом:

$$\xi_1(t) = \eta_1(t), \quad \xi_r(t) = \eta_r(t) - \sum_{k=1}^{r-1} \int_0^t \alpha_{rk}(s) d\xi_k(s), \quad r > 1,$$

$$\alpha_{rk}(s) = \frac{d\langle \eta_r, \xi_k \rangle}{d\langle \xi_k \rangle_s}.$$

Построенная последовательность $\{\xi_k(t)\}$ обладает следующим свойством: для всякого мартингала $\eta(t) \in M_2^0$ справедливо представление

$$\eta(t) = \sum_{k=1}^{\infty} \int_0^t \alpha_k(s) d\xi_k(s), \quad \alpha_k(s) = \frac{d\langle \eta, \xi_k \rangle_s}{d\langle \xi_k \rangle_s}.$$

Всякую последовательность, обладающую этим свойством, будем называть базисом в M_2^0 .

ТЕОРЕМА 4. Пусть $\delta(t)$ — существенное время, $\{\xi_k(t)\}$ и $\{\tilde{\xi}_k(t)\}$ — два базиса в M_2^0 . Тогда если

$$\nu(t) = \sum_k I_{\{d\langle \xi_k \rangle_t / d\delta(t) > 0\}}, \quad \tilde{\nu}(t) = \sum_k I_{\{d\langle \tilde{\xi}_k \rangle_t / d\delta(t) > 0\}},$$

то

$$P\left\{\int I_{\{\nu(s) \neq \tilde{\nu}(s)\}} d\delta(s) = 0\right\} = 1.$$

Таким образом существует не зависящая от выбора базиса такая \mathcal{P} -измеримая функция $r(t)$, что

$$P\{\nu(t) = r(t)\} = 1,$$

$r(t)$ называется рангом потока (\mathcal{F}_t) в момент t . А величина

$$r = \inf \left\{ k: P \left\{ \int I_{\{\nu(s) > k\}} d\delta(s) = 0 \right\} = 1 \right\}$$

(считаем $\inf \emptyset = +\infty$) называется рангом потока (\mathcal{F}_t) . Ранг потока — это размерность минимального базиса в M_2^0 , это вытекает из следующего утверждения.

ТЕОРЕМА 5. Пусть $\delta(t)$ существенное время, r — ранг (\mathcal{F}_t) . Существует последовательность \mathcal{P} -измеримых множеств $A_1 \supset A_2 \supset \dots \supset A_r$ (эта последовательность бесконечна при $r = +\infty$) и базис $\{\xi_k(t), k < r+1\}$ такие, что

$$\langle \xi_i \rangle_t = \int_0^t I_{A_i}(s) d\delta(s).$$

5. Примеры. Мартингалы из $M_2^1 \oplus M_2^2$ имеют довольно простую структуру, хотя для M_2^1 мы укажем далее некоторый определяющий инвариант. Сейчас наша цель показать, что ранг процесса может быть различным, причем процесс при этом может быть довольно регулярным процессом в R .

ПРИМЕР 1. Пусть $x(t, \omega)$ — процесс в R^m , его компоненты $x_1(t, \omega), \dots, x_m(t, \omega)$ — независимые винеровские процессы. Покажем, что его ранг $r = m$. Так как для всякого мартингала, согласованного с потоком (\mathcal{F}_t) , порожденным $x(t, \omega)$, справедливо представление

$$\eta(t) = \sum_{k=1}^m \int_0^t \varphi_k(s) dx_k(s, \omega),$$

то $r \neq m$. Если бы существовал базис из $r < m$ мартингалов $\eta_1(t), \dots, \eta_r(t)$ и

$$x_k(t, \omega) = \sum_{i=1}^r \int_0^t \alpha_{ki}(s) d\eta_i(s),$$

то

$$\delta_{lk}t = \sum_{i=1}^r \int_0^t \alpha_{ki}(s) \alpha_{li}(s) d\langle \eta_i \rangle_s, \quad \delta_{lk} = \sum_{i=1}^r \alpha_{li} \alpha_{ki}(t) \frac{d\langle \eta_i \rangle_t}{dt},$$

$l, k \leq m$, что невозможно (ранг матрицы $\|\alpha_{ki}(d\langle \eta_i \rangle/ds)^{1/2}\|$ не превосходит $r \leq m$).

ПРИМЕР 2. Пусть $\{w_k(t)\}$ последовательность независимых винеровских процессов. Положим

$$\eta_{n,n+1}(t) = \int_0^t (\arctg w_{n+1}(s) + \pi/2) dw_n(s),$$

$$\eta_{n,m}(t) = \int_0^t (\arctg \eta_{n+1,m}(s) + \pi/2) dw_n(s), \quad m > n + 1.$$

Поток $(\mathcal{F}_t^{n,m})$, порождаемый мартингалом $\eta_{n,m}(t)$, совпадает с потоком, порождаемым винеровскими процессами $w_n(t), w_{n+1}(t), \dots, w_m(t)$, и его ранг равен $m - n + 1$. Легко убедиться, что существует предел в среднеквадратической сходимости $\lim_{m \rightarrow \infty} \eta_{n,m}(t) = \eta_n(t)$ и поток \mathcal{F}_t^n , порожденным предельным процессом, совпадает с потоком, порожденным винеровскими процессами $w_n(t), w_{n+1}(t), \dots$. Он имеет ранг $+\infty$. В то же время процессы $\eta_{n,m}(t)$ и $\eta_n(t)$ — одномерные непрерывные мартингалы с дифференцируемыми характеристиками.

6. Квазинепрерывный поток (\mathcal{F}_t) . Случайная замена времени. Поток называется квазинепрерывным, если M_2^2 содержит только 0. В этом случае характеристики всех мартингалов из M_2 непрерывны. Пусть $\delta(t)$ существенное время и $\delta(t) \uparrow +\infty$ при $t \uparrow +\infty$. Положим $\tau_t = \delta^{-1}(t)$, $\tilde{x}(t, \omega) = x(\tau_t, \omega)$, $(\tilde{\mathcal{F}}_t) = (\mathcal{F}_{\tau_t})$, $\tilde{M}_2, \tilde{M}_2^0, \tilde{M}_2^1, \tilde{M}_2^2$ отвечают потоку $(\tilde{\mathcal{F}}_t)$.

ТЕОРЕМА 6. $\tilde{M}_2 = \{\eta(\tau_t), \eta \in M_2\}$, $\tilde{M}_2^i = \{\eta(\tau_t), \eta \in M_2^i\}$, $i = 0, 1, 2$. Заметим, что процесс $\tilde{x}(t, \omega)$ не обязательно порождает $(\tilde{\mathcal{F}}_t)$, он вообще может стать неслучайным.

Поток называется абсолютно непрерывным, если он непрерывен и существенное время абсолютно непрерывно (дифференцируемо по t). Приведенная выше замена времени приводит к тому, что поток $(\tilde{\mathcal{F}}_t)$ абсолютно непрерывен.

ТЕОРЕМА 7. Для абсолютно непрерывного потока (\mathcal{F}_t)

(1) существует последовательность винеровских процессов $\{w_k(t)\}$ и пуассоновская мера $\pi(dx \times dt)$ на R_+^2 , для которой $E\pi(ds \times dx) = dx dt$, такие, что всякий мартингал $\eta(t) \in M_2$ представим в виде

$$\eta(t) = \sum_{k=1}^{\infty} \int_0^t \alpha_k(s) dw_k(s) + \int_0^t \int_{R_+} \gamma(s, x) [\pi(ds \times dx) - ds dx], \quad (2)$$

где $\alpha_k(s)$ и $\gamma(s, x)$ — некоторые функции, измеримые относительно \mathcal{P} и $\mathcal{P} \otimes \mathcal{B}_{R_+}$ соответственно, для которых при всех $t > 0$

$$P \left\{ \sum \int_0^t \alpha_k^2(s) ds + \int_0^t \int_{R_+} \gamma^2(s, x) ds dx < \infty \right\};$$

(2) существуют такие \mathcal{P} -измеримые функции $r(t)$ и $\lambda(t)$, принимающие значения: первая — $0, 1, 2, \dots, +\infty$, вторая — из R_+ или $+\infty$ такие, что \mathcal{F}_t совпадает с потоком, порожденным процессами

$$\left\{ \int_0^t I_{\{r(s) \geq k\}} dw_k(s), \quad k = 1, 2, \dots; \quad \int_0^t \int_{R_+} I_{\{x \leq \lambda(t)\}} \frac{1}{1+x^2} \pi(ds \times dx) \right\};$$

$r(t)$ — ранг потока. Функция $\lambda(t)$ имеет следующий смысл. Пусть $\nu(t) - \nu(s)$ при $s < t$ — число скачков всех мартингалов из M_2^1 на отрезке $[s, t]$, тогда $\int_s^t \lambda(u) du$ есть компенсатор $\nu(t) - \nu(s)$; $\lambda(t)$ называется шириной пуассоновского спектра (\mathcal{F}_t) . Это тоже величина, инвариантная при преобразованиях процесса, сохраняющих поток.

7. Неупреждающие преобразования и стохастические уравнения.

Пусть $\{\eta_n(t)\}$ — такая последовательность мартингалов из M_2 , что для всех t линейная оболочка величин $\{\eta_n(t), n \geq 1\}$ плотна в H_t . Будем предполагать, что поток абсолютно непрерывен. Тогда для каждого мартингала $\eta_n(t)$ можно записать формулу (2) с функциями $\alpha_{nk}(s)$ и $\gamma_n(s, x)$. При фиксированном x — это \mathcal{F}_s -измеримые случайные величины, входящие в H_s . Значит они могут быть выражены как некоторая линейная функция от $\{\eta_n(s), n = 1, 2, \dots\}$. Более того, поскольку они предсказуемы, а предсказуемая проекция $\eta_n(s)$ есть $\eta_n(s-)$, то их можно рассматривать как линейные функции от $\eta_n(s-)$. Будем формально записывать эти линейные функции в виде бесконечных

линейных комбинаций. (На самом деле это пределы конечных линейных комбинаций.) Если

$$\begin{aligned}\alpha_{nk}(s) &= \sum a_{km}^n(s) \eta_m(s-), \\ \gamma_n(s) &= \sum g_m^n(s, x) \eta_m(s-)\end{aligned}$$

(функции $a_{km}^n(s)$ и $g_m^n(s, x)$ уже неслучайны), то для последовательности $\{\eta_n(t)\}$ получаем следующую систему линейных стохастических дифференциальных уравнений

$$\begin{aligned}d\eta_n(t) &= \sum_{k=1}^{\infty} \left(\sum_{m=1}^{\infty} a_{km}^n(s) \eta_m(s-) \right) dw_k(s) \\ &+ \int_{R_+} \sum_{m=1}^{\infty} g_m^n(s, x) \eta_m(s-) [\pi(ds \times dx) - ds dx].\end{aligned}\quad (3)$$

Уравнение (3) можно рассматривать как линейное уравнение в R^∞ с неограниченными операторными коэффициентами. Уточнение смысла уравнений (3), а также исследование их решений (в частности, условий единственности) — задача теории линейных стохастических дифференциальных уравнений.

Поскольку процесс $\eta_n(t)$ \mathcal{F}_t -измерим, а \mathcal{F}_t порождено значениями процесса $x(s, \omega)$ при $s \leq t$, то существует такая $\mathcal{B}_{R_+} \otimes \mathcal{B}^{R_+}$ -измеримая функция $f_n(t, x(\cdot))$ из $R_+ \times X^{R_+}$ в R , что

$$\eta_n(t) = f_n(t, x(\cdot, \omega)),$$

при этом функция f_n неупреждающая: если $y_1(s) + y_2(s)$ при $s \leq t$, $y_i(\cdot) \in X^{R_+}$, то $f_n(t, y_1(\cdot)) = f_n(t, y_2(\cdot))$. Отображение процесса $x(t, \omega) \rightarrow \{f_n(t, x(\cdot, \omega)), n = 1, 2, \dots\}$ в R^∞ сохраняет поток (\mathcal{F}_t) и поэтому оно обратимо в обобщенном смысле. Это и есть то обратимое неупреждающее преобразование в R^∞ , которое после случайной замены времени приводит квазинепрерывный процесс к процессу, удовлетворяющему системе линейных стохастических дифференциальных уравнений.

8. Некоторые выводы и задачи. 1. Вполне понятно строение квазинепрерывного потока. Он определяется такими тремя характеристиками: (а) существенным временем $\delta(t)$, которое определяется с точностью до эквивалентности для мер, порождаемых возрастающими непрерывными функциями; (б) рангом процесса $r(t)$, определенным однозначно почти всюду относительно существенного времени; (в) шириной пуассоновского спектра $\lambda(t)$, которая определяется однозначно почти всюду относительно существенного времени, если это время фиксировано.

Заметим, что функции $r(t)$ и $\lambda(t)$, определяющие поток при $\delta(t) = t$, должны быть предсказуемы относительно определяемого потока. Это накладывает на них определенные ограничения, которые следует изучить.

2. Выясняется особая роль линейных стохастических дифференциальных уравнений с неограниченными операторными коэффициентами. Изучение таких уравнений проводилось в рамках теории линейных систем при весьма жестких ограничениях, оказывается имеет смысл рассматривать такие системы без всяких ограничений.

3. Рассмотрены только некоторые преобразования процессов и потоков σ -алгебр. Весьма интересен вопрос о необратимых преобразованиях процессов (например, с помощью преобразований фазового пространства), при которых сохраняются потоки σ -алгебр.

4. Было бы полезно иметь способ доказательства квазинепрерывности потока, порожденного случайным процессом без рассмотрения пространства мартингалов, а используя лишь его конечномерные распределения. (Это возможно, например, для марковских процессов.)

5. Интерес представляет изучение общих потоков (с предсказуемыми разрывами). Здесь можно было бы использовать операцию предельного перехода для потоков с конечным числом предсказуемых разрывов.

ЛИТЕРАТУРА

1. А. В. Скороход, *О локальном строении непрерывных марковских процессов*, Теория вероятностей и ее применения 1 № 13 (1966), 381–423.
2. —, *Операторные стохастические дифференциальные уравнения и стохастические полугруппы*, Успехи мат. наук 37 № 6 (1982), 157–183.
3. И. И. Гихман и А. В. Скороход, *Стохастические дифференциальные уравнения и их приложения*, Наук. думка, Киев, 1982.
4. А. В. Скороход, *Стохастические уравнения для сложных систем*, Наука, Москва, 1983.

Институт математики, Академии наук украинской ССР, Киев-4, СССР

Algorithms for Solving Equations

STEVE SMALE

The main goal of this work is an attempt to understand the efficiency of algorithms for solving systems of equations. For example, consider a map f from complex Cartesian space \mathbf{C}^n to \mathbf{C}^n given by polynomials. How fast can one find good approximations to one or all of the zeros of f ? However, for clarity of exposition and development we will usually swing from one extreme to another about this example, proving theorems on one hand for a single complex polynomial and on the other hand for an analytic map $f: \mathbf{F} \rightarrow \mathbf{F}$ from one Banach space to another (both real or both complex).

This subject should undoubtedly be considered as numerical analysis. However, my point of view has been especially influenced by complexity theory of computer science. Thus the emphasis is on the algorithms themselves and a global study of their speed. Hence less emphasis is given to the results obtained by algorithms and the asymptotic criteria of efficiency that are often found in numerical analysis literature.

This global study, or complexity theory, gives a more systematic, more abstract, more theoretical flavor to our approach, and thus we are less concerned about producing immediately faster methods for problem solving. A basic understanding of the tried and effective is sought. The global study of the algorithm forces the introduction of topology and geometry into the subject.

If any algorithm has proved itself for the problem of nonlinear systems, it is Newton's method and its many modifications. The Greeks essentially used it for finding square roots; and for that purpose it is still a top method today. On the other hand, for nonlinear functional equations framed in a Banach space setting, Newton's method finds a central place.

In the account here, the one theme is Newton's method. We will prove theorems about it for one variable (dealing with the fundamental theorem of algebra) and for maps of Banach spaces; we will approximate Dantzig's [8] simplex method for the linear programming problem by Newton's method.

Supported in part by National Science Foundation grants and by the Mathematical Sciences Research Institute.

The contributions and point of view of Kantorovich dominate the modern treatment of Newton's method. For background see Berger [2], Henrici [19], Kantorovich-Akilov [23], and Ostrowski [38]. This treatment has great beauty derived from its simplicity, its generality, and its weakness of its hypotheses. The assumptions in a Banach space context are first and second derivative bounds on a domain of a map.

What we do is to start afresh with analytic maps and *derive estimates at one point*.

In the execution of an iterative algorithm, one finds oneself at some point in the domain space and is able to compute derivatives at that point. From that information, it is important to make a judgment for the next step, or perhaps even terminate the algorithm. These considerations motivated our paper, [52], referred to subsequently as [point estimates] or just [P.E.]. Since the results of [P.E.] are used frequently below, we review a bit of that theory.

One aspect of a theoretical study of problems where only approximate solutions are possible, is the omnipresence of a small parameter $\varepsilon > 0$, which measures the distance to a solution. However, for a well-conditioned solution of $f(\zeta) = 0$, f a complex polynomial, one can dispense with such arbitrariness using the notion of an approximate zero.

One variable, Newton's method for solving $f(\zeta) = 0$, with starting point $z_0 \in \mathbf{C}$, is the sequence $z_k = z_{k-1} - f(z_{k-1})/f'(z_{k-1})$ defined inductively. The point z_0 will be called an *approximate zero* if

$$|z_k - z_{k-1}| \leq \left(\frac{1}{2}\right)^{2^{k-1}-1} |z_1 - z_0|, \quad k = 1, 2, 3, \dots$$

Thus if $k = 5$, for example, the coefficient is already $(\frac{1}{2})^{15}$. In machine computations one sees this phenomenon with doubling of precision at each step. However, it would be satisfying to have a test that one could make with information at z_0 to insure this estimate for all k . That motivates the introduction of our invariant $\alpha(z, f)$ and Theorems A and B as follows.

Let $\beta(z, f)$ be the length of the "Newton vector"

$$\beta(z, f) = \left| \frac{f(z)}{f'(z)} \right| \quad \text{and} \quad \gamma(z, f) = \max_{k \geq 1} \left| \frac{f^{(k)}(z)}{k! f'(z)} \right|^{1/(k-1)}.$$

Here $f^k(z)$ is the k th derivative. Then the invariant $\alpha(z, f)$ is defined as the product $\beta(z, f)\gamma(z, f)$ or $\beta(z)\gamma(z)$.

THEOREM A (SPECIAL CASE). *There is a constant α_0 equal to approximately .130707 such that if $\alpha(z, f) < \alpha_0$, then z is an approximate zero of f .*

The general case of Theorem A has the same statement, but f and the definition of approximate zero are generalized as follows.

Now $f: \mathbf{E} \rightarrow \mathbf{F}$ is an analytic map from one Banach space to another (e.g., $\mathbf{E} = \mathbf{C}^n = \mathbf{F}$).

Newton's method starting from $z_0 \in E$ is given by

$$z_k = z_{k-1} - Df(z_{k-1})^{-1}f(z_{k-1})$$

and z_0 is an approximate zero of f if $\|z_k - z_{k-1}\| < (\frac{1}{2})^{2^{k-1}-1} \|z_1 - z_0\|$, $k = 1, 2, \dots$. Let

$$\beta(z, f) = \|Df(z)^{-1}f(z)\| \quad \text{and} \quad \gamma(z, f) = \sup_k \left\| Df(z)^{-1} \frac{D^k f(z)}{k!} \right\|^{1/(k-1)}$$

where $D^k f(z)$ is the k th derivative of f as a k -linear functional (see Dieudonné [10] or Lang [30] for the calculus). In case $Df(z)^{-1}$ is not defined, we make α, β , and γ infinite. Now Theorem A makes sense for analytic $f: \mathbf{E} \rightarrow \mathbf{F}$ and is true; this is proved in [P.E.].

M. Kim [27] independently has a proof of Theorem A with one variable and $\alpha_0 = \frac{1}{54}$. I wrote in [P.E.] that theorems similar to ours could probably be deduced with the Kantorovich theory as a starting point. Subsequently, H. Royden [42] showed that that indeed was the case with a slight improvement of α_0 in Theorem A. Moreover, recently J. Curry [7] has extended the one-dimensional version of Theorem A to higher order generalizations of Newton's method.

One might object to the feature of Theorem A that requires knowledge of *all* derivatives for the γ factor of α . In fact only the first derivative is necessary, together with some crude information on f . To make this precise, define the function $\varphi_d: \mathbf{R} \rightarrow \mathbf{R}$ by $\varphi_d(r) = \sum_{i=0}^d r^i$. Then φ'_d is the derivative.

If f is a polynomial described by $f(z) = \sum_{i=0}^d a_i z^i$, let $\|f\| = \max_i |a_i|$.

THEOREM B (SPECIAL CASE). *For each polynomial f and each $z \in \mathbf{C}$,*

$$\gamma(z, f) \leq \frac{\|f\|}{|f'(z)|} \frac{\varphi'_d(|z|)^2}{\varphi_d(|z|)}.$$

For the general case, let $f: \mathbf{E} \rightarrow \mathbf{F}$ be an analytic map of Banach spaces which can be expressed as $f(z) = \sum_{k=0}^{\infty} a_k z^k$ where the a_k are symmetric k -linear maps and $a_k z^k$ is a_k evaluated at the k -tuple (z, \dots, z) of elements of \mathbf{E} . Let $\|a_k\|$ be the usual norm as for example in Lang [30]. Then define $\|f\| = \sup_k \|a_k\|$.

THEOREM B. *With $f: \mathbf{E} \rightarrow \mathbf{F}$, $z \in \mathbf{E}$ as above*

$$\gamma(z, f) \leq \|f\| \|Df(z)^{-1}\| \frac{\varphi'_d(\|z\|)^2}{\varphi_d(\|z\|)}.$$

Theorem B is announced in [P.E.] and proved in §1 below. It is used in our proofs of the theorems below on the Tractability of the Random Algorithm and Average Area of Approximate Zeros.

By providing a termination test, Theorem A allows us to introduce a simple Random Algorithm for polynomial zero finding. This goes as follows: Given a complex polynomial f ,

- (1) Choose $z \in \mathbf{C}$, $|z| < 3$, at random.
- (2) Is $\alpha(z, f) < \alpha_0$? If yes, terminate (or apply Newton's method, say 5 times, then terminate). If no, go to (1).

It is natural to ask, for a given f : How many steps will the random algorithm take on the average? For this to make sense let Ω be the set of all possible

sequences $\mathbf{Z} = (z_1, z_2, z_3, \dots)$, $|z_k| < 3$. Give the disk of radius 3 normalized Lebesgue measure and endow Ω with the infinite product measure as a probability measure, as in Shub-Smale [44].

Define the function $\sigma: \Omega \rightarrow \mathbb{Z}^+$ by $\sigma(Z)$, which is the first σ such that $\alpha(z_\sigma, f) < \alpha_0$. Then the *average number of steps* needed to terminate the random algorithm is defined as

$$\sigma_f = \text{average}_{Z \in \Omega} \sigma(Z).$$

THEOREM (THE TRACTABILITY OF THE RANDOM ALGORITHM). *The set of polynomials f in $P_d(1)$, such that average number of steps is larger than σ , has measure less than cd^5/σ for an appropriate constant c .*

Of course one hopes for a better bound eventually. This theorem is proved in §2.

Here $P_d(1)$ is the set of all complex polynomials $f(z) = \sum_0^d a_i z^i$ with $a_d = 1$ and $|a_i| \leq 1$. We have imposed uniform probability measure on $P_d(1)$ (i.e., normalized Lebesgue measure) using the fact that $P_d(1)$ is a bounded set in \mathbb{C}^d . The set of $f \in P_d(1)$ with $\sigma_f = \infty$ has measure 0. These f are the ones where zero finding is an ill-posed problem in some reasonable sense.

For the next result, let $z_0 \in \mathbb{C}$, f a complex polynomial, and consider

$$|z_n - \zeta| < \left(\frac{1}{2}\right)^{2^n-1} |z_0 - \zeta| \quad (*)$$

where $f(\zeta) = 0$ and $z_n = z_{n-1} - f(z_{n-1})/f'(z_{n-1})$. Thus $(*)$ is another form of a superconvergent property independent of a small parameter. Let Ω_f be the set of $|z_0| < 1$ satisfying $(*)$, and let A_f be its area.

THEOREM (AVERAGE AREA OF APPROXIMATE ZEROS).

$$\text{average}_{f \in P_d(1)} A_f > c > 0$$

where c is independent of d .

This result is equally valid if f is averaged over the set $\{f | f(z) = \sum_{i=0}^d a_i z^i, |a_i| \leq 1, i = 0, 1, \dots, d\}$. The proof is given in §3.

Points z_0 satisfying $(*)$ are called approximate zeros of the second kind. Generally if $f: \mathbf{E} \rightarrow \mathbf{F}$ is an analytic map, $z_0 \in \mathbf{E}$, $z_n = z_{n-1} - Df(z_{n-1})^{-1}(F(z_{n-1}))$,

$$\|z_n - \zeta\| \leq \left(\frac{1}{2}\right)^{2^n-1} \|z_0 - \zeta\|, \quad f(\zeta) = 0,$$

then z_0 will be called an *approximate zero of the second kind*. For the proof of the previous theorem we use Theorem C which was proved in [P.E.].

THEOREM C. *Suppose that $f: \mathbf{E} \rightarrow \mathbf{F}$ is analytic, $\zeta \in \mathbf{E}$, $f(\zeta) = 0$, and $z \in \mathbf{E}$ satisfies*

$$\|z - \zeta\| < \frac{3 - \sqrt{7}}{2} \frac{1}{\gamma(\zeta, f)}.$$

Then z is an approximate zero of the second kind.

Next, consider an analytic map $f: \mathbf{E} \rightarrow \mathbf{F}$ from one Banach space to another. We will study an algorithm based on Newton's method for approximating solutions of $f(\zeta) = 0$. Complexity aspects of this algorithm will be proved.

We suppose f given as above and $z_0 \in \mathbf{E}$. It is important for our purposes to understand the *lifted path* $\sigma = \sigma(z_0, f) \subset \mathbf{E}$, defined as follows. If the derivative $Df(z_0): \mathbf{E} \rightarrow \mathbf{F}$ is an isomorphism, let $f_{z_0}^{-1}$ be the local inverse of f defined from a neighborhood of $f(z_0)$ to \mathbf{E} and taking $f(z_0)$ to z_0 . For nonnegative t_0 sufficiently close to one, $f_{z_0}^{-1}$ maps $\{tf(z_0) | t_0 < t \leq 1\}$ into \mathbf{E} .

It may happen that $f_{z_0}^{-1}$ extends to a map of the ray $\{tf(z_0) | 0 \leq t \leq 1\} \rightarrow \mathbf{E}$, remaining an inverse to f and such that the derivative along this ray is always an isomorphism. In this case we say that $\sigma = \sigma(z_0, f)$ is defined and is the set $f_{z_0}^{-1}\{tf(z_0) | 0 \leq t \leq 1\}$.

If f is a complex polynomial of one variable, the curve σ is discussed and plays a major role in Smale [50, 51] and Shub-Smale [43, 44].

Moreover, $\sigma(z_0, f)$ is defined generically for systems $f: \mathbf{C}^n \rightarrow \mathbf{C}^n$ given by polynomials.

With z_0, f as above, we define an invariant $M(z_0, f)$ as follows. If $\sigma(z_0, f)$ is not defined, let $M(z_0, f) = \infty$. Otherwise, $M(z_0, f) = \max_{z \in \sigma} \alpha(z, f) / \|f(z)\|$.

Our algorithm for solving $f(\zeta) = 0$ is contained in the following theorem.

THEOREM (THE SPEED OF A GLOBAL NEWTON METHOD). *There exist (small) positive constants, c real, l an integer with this property. Let $f: \mathbf{E} \rightarrow \mathbf{F}$ be analytic, $z_0 \in \mathbf{E}$, with $M(z_0, f) < \infty$. Suppose n is an integer,*

$$n > c\|f(z_0)\|M(z_0, f), \quad \Delta = 1/n.$$

Let $w_i = (1 - i\Delta)f(z_0)$, $i = 0, \dots, n$. Then, inductively, $z_i = N_{f-w_i}^l(z_{i-1})$ is well defined and z_n is an approximate zero of f .

Here $N_{f-w_i}(z_{i-1})$ is Newton's method for solving $f(\zeta) - w_i = 0$, applied to z_{i-1} , and $N_{f-w_i}^l(z_{i-1})$ is the same applied iteratively l times. The constants c and l are about 4. We give the proof in §4.

The algorithm in the above theorem is a version of the Global Newton in Smale [47], Hirsch-Smale [20], Keller [26], Garcia-Gould [15], Abadie-Guerrero [1]. Related results can be found in Kung [28], and the book of Dejon-Henrici [29] (especially the article by Dejon-Nickel in the last volume). Surely the theorem has roots in the one-dimensional case of Shub-Smale [43, 44] and Smale [50, 51].

The ideas of this proof extend to deal with other situations, for example, homotopy problems as in Keller [25] and Chow-Mallet-Paret-Yorke [4]. Moreover, if a problem of zero finding is ill-posed, the corresponding residual problem $|f(z)| < \varepsilon$ could be well-posed and the complexity theory of the above theorem applies (again similar to Smale [50, 51] and Shub-Smale [43, 44]).

Both Newton's method and piecewise linear algorithms can be thought of as path following algorithms. This is a theme in Eaves-Scarf [11], Smale [47], and Hirsch-Smale [20]. One can interpret Dantzig's [8] self-dual variant of the

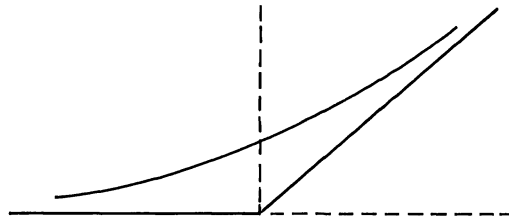


FIGURE 1

simplex method as a path following method (Smale [48]) in this way. Thus a relation between the simplex method of linear programming and Newton's method, is no surprise. In fact, the relationship turns out to be very simple and natural in the framework of the linear complementary problem or LCP.

The LCP starts with data (M, q) where M is an $N \times N$ matrix and $q \in \mathbf{R}^N$.

Define the functions $x^\pm = (x \pm |x|)/2$ for a real variable x (more formally one could write $x^+(x) = (x + |x|)/2$ etc.). Let $\Phi_M: \mathbf{R}^N \rightarrow \mathbf{R}^N$ be the map $\Phi_M(x) = x^+ + Mx^-$ where $x^\pm = (x_1^\pm, \dots, x_N^\pm)$. The LCP is: given (M, q) solve for x ;

$$\Phi_M(x) = q. \quad (\text{LCP})$$

For background, see Cottle-Dantzig [6], Smale [48, 49], and the cited references.

We are especially concerned here with how the LCP can be considered as containing the linear programming problem, or LPP, as a special case. The LPP with data (A, b, c) , A an $m \times n$ matrix, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$ is:

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \ c \cdot x \quad \text{subject to} \ x \geq 0, \ Ax \geq b. \quad (\text{LPP})$$

Here $c \cdot x = \langle c, x \rangle$ is the inner product.

Then the LPP data (A, b, c) generate LCP data (M, q) by $N = n + m$, $q = (c, -b)$, and $M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}$. It will be assumed that M has this special form.

The LPP has a solution if and only if the corresponding LCP does, and in that case the solutions correspond in a natural way.

We propose here an analytic approximation to Φ_M , which at the same time desingularizes the LPP and makes analytic methods—Newton's method, in particular—available for its solutions.

In this way, Dantzig's [8] self-dual algorithm, a version of his simplex method, can be approximated by Newton's method.

We start with an approximation of the functions of one variable, x^\pm . Let $\varphi_a^\pm: \mathbf{R} \rightarrow \mathbf{R}$ be defined by

$$\varphi_a^\pm(x) = \frac{x \pm (x^2 + a^2)^{1/2}}{2} \quad \text{for } a \geq 0.$$

The graph of φ_a^+ for $a > 0$ is a hyperbola (see Figure 1).

As $a \rightarrow 0$, φ_a^\pm approaches x^\pm uniformly, and as $x \rightarrow \pm\infty$ the approximation only improves.

Next let $\Phi_a: \mathbf{R}^N \rightarrow \mathbf{R}^N$ be the map

$$\Phi_a(x) = \Phi_a^+(x) + M\Phi_a^-(x)$$

where

$$\Phi_a^\pm(x) = (\varphi_a^\pm(x_1), \dots, \varphi_a^\pm(x_N)).$$

From the above observations, Φ_a tends to Φ_M uniformly on all of \mathbf{R}^N as $a \rightarrow 0$; and for each a , $\Phi_a(x)$ tends to $\Phi_M(x)$ as $\|x\| \rightarrow \infty$.

Our next goal is to give some analysis of Φ_a for $a > 0$. Define $\mathcal{U}_M \subset \mathbf{R}^N$ as the cone of all positive linear combinations of the column vectors of $-M$ and the coordinate vectors e_1, \dots, e_N .

A diffeomorphism is a differentiable map with a differentiable inverse. It is analytic if, in a neighborhood of each point, it can be defined by a convergent power series.

THEOREM (ON THE REGULARIZATION OF THE LPP). *For each $a > 0$, the map $\Phi_a: \mathbf{R}^N \rightarrow \mathbf{R}^N$ has image \mathcal{U}_M and is an analytic diffeomorphism $\Phi_a: \mathbf{R}^N \rightarrow \mathcal{U}_M$.*

This theorem is proved in §5.

REMARK 1. Actually the proof does not need the special form of M . The proof and the theorem are valid for any square matrix satisfying $\langle Mx, x \rangle \geq 0$. The same is true for the proposition in §5.

REMARK 2. As a consequence of the theorem, given M for any $q \in \mathcal{U}_M$, and any $a > 0$, there is a unique solution of the problem $\Phi_M(x) = q$. This solution depends analytically on the data (A, b, c) of the LPP and a , as long as $q \in \mathcal{U}_M$ and $a > 0$.

Say that the LPP with Data (A, b, c) is “extremely ill-posed” if $q \in \partial\mathcal{U}_M$ ($\partial\mathcal{U}_M = \overline{\mathcal{U}_M} - \mathcal{U}_M$). Here is some justification for our terminology. The LPP has a solution if and only if $q \in \overline{\mathcal{U}_M}$ (see Cottle [5]). Thus if $q \in \partial\mathcal{U}$, a small perturbation of the data means that the problem has no solution. The above theorem implies that all LPP which are not extremely ill-posed can be desingularized at once, by taking any $a > 0$.

Newton’s method for $\Phi_a - q$ follows the curves $\Phi_a^{-1}(q\hat{q})$ where $q\hat{q}$ is the segment from an initial value \hat{q} to q . Dantzig’s method follows the curve $\Phi_a^{-1}(q\hat{q})$ for $a = 0$. Thus as $a \rightarrow 0$, Newton’s method, with an appropriate step size, tends to the Dantzig Algorithm.

It is convenient to introduce a map $\hat{\Phi}_a: \mathbf{R}^N \rightarrow \mathbf{R}^N$ by $\hat{\Phi}_a(x) = \Phi_a(x) - \Phi_a(0)$. Then as a goes to zero, $\hat{\Phi}_a$ approaches the LPP. Moreover, as is shown in §5, Theorem A can be applied to solve $\hat{\Phi}_a(x) = q$ for all a such that $a > c\|q\|$. Then the original LPP problem can be obtained by reducing a to zero.

Related in one way or another to the previous result is the work of Mangasarian [31], Wierzbicki [53], Karmarkar [24], Gill-Murray-Saunders-Tomlin-Wright [17], Blum [3], Renegar [41], and Megiddo-Shub [36].

More generally, there are many other papers related to the subject of this paper. In particular, in the area of information based complexity, one can see Woźniakowski [55] for a recent survey. Smale [51] has many other references. Papers not listed in [51], but which have interesting related results, include Pan [39], Wong [54], Gao [13, 14], and Wright [56].

I would like to express my thanks to a number of mathematicians for their help in this paper. In particular, a conversation with Dick Cottle on the LCP was useful. Lenore Blum's MSRI talk on a condition number for linear programming via the LCP helped put the LCP back in my mind. Her comments and those of Jim Curry and Feng Gao have been generally useful to me.

Especially important through all of this has been the work of, and conversations with, Jim Renegar and Mike Shub. That contribution from Mike Shub, to me, has persisted over many years indeed.

1. Suppose α is defined as above and $\varphi_d(r) = \sum_{j=0}^d r^j$. Toward the proof of Theorem B, we show:

THEOREM 1. $\alpha(r, \varphi_d) \leq 1$ for all $r \geq 0$.

PROOF. First consider the limit $d = \infty$, so that $\varphi_\infty(r) = \varphi(r) = 1/(1-r)$. Then

$$\frac{\varphi^{(k)}(r)}{k!} = \left(\frac{1}{1-r} \right)^{k+1}$$

and from the definition of α it follows that $\alpha(r, \varphi) \equiv 1$.

Now define

$$Q_p(d) = \sum_{j=p}^d \binom{j}{p}$$

where p, d, j are nonnegative integers with $d \geq p$ (but recall the binomial coefficient $\binom{j}{p} = 0$ if $j < p$).

LEMMA 1.

$$Q_p(d) = \frac{(d+1)d(d-1)\cdots(d-p+1)}{(p+1)!}.$$

PROOF OF LEMMA 1. Note that $Q_p(d)$ is a polynomial in d of degree $p+1$. Then it is sufficient to check the formula of the lemma at $p+2$ values of d . Note then $Q_p(d) = 0$ for $d = 0, 1, \dots, p-1$, $Q_p(p) = 1$, and $Q_p(p+1) = p+2$. Q.E.D.

LEMMA 2. For $0 \leq m \leq d$, $(m+1)Q_p(d) \geq (d+1)Q_p(m)$.

Use Lemma 1 to write out the inequality of Lemma 2 as

$$\frac{(m+1)(d+1)\cdots(d-p+1)}{p!} \geq \frac{(d+1)(m+1)m\cdots(m-p+1)}{p!}.$$

Lemma 2 is immediate.

For $k = 0, 1, 2, \dots, d$ introduce

$$\psi_k(x) = \frac{\varphi^{(k)}(x)}{k!} = \sum_{l=0}^{d-k} \binom{l+k}{k} x^l = \sum_{j=0}^d \binom{j}{k} x^{j-k}.$$

In these terms the estimates of the theorem is $\psi_0^{k-1}\psi_k \leq \psi_1^k$, $k = 1, 2, \dots$. For $k = 1$, this is trivial. We proceed to prove the theorem by induction on k . Thus it is sufficient to show $\psi_0\psi_k \leq \psi_1\psi_{k-1}$, $k = 2, 3, \dots$, $x > 0$.

Write out this last estimate and multiply both sides by x^k to obtain

$$\sum_0^d x^i \sum_0^d \binom{j}{k} x^j \leq \sum_0^d i x^i \sum_0^d \binom{j}{k-1} x^j.$$

The product on the left can be written in the form $\sum_{m=0}^{2d} c_m x^m$ while that on the right has the form $\sum_{m=0}^{2d} \sigma_m x^m$. It is sufficient to show that $c_m \leq \sigma_m$, all m .

LEMMA 3. $c_m = \sigma_m$ for $m \leq d$.

This is a consequence of the fact that $\alpha(r, \varphi) \equiv 1$. For $m \leq d$, we might as well have $d = \infty$.

Next we evaluate the c_m, σ_m . If $m \leq d$, then

$$c_m = \sum_{i=0}^m \binom{j}{k} = Q_k(m), \quad \sigma_m = \sum_{j=0}^m (m-j) \binom{j}{k-1}.$$

If $d < m \leq 2d$, then

$$c_m = \sum_{j=m-d}^d \binom{j}{k}, \quad \sigma_m = \sum_{j=m-d}^d (m-j) \binom{j}{k-1}.$$

Define $P_{k-1}(q) = \sum_{j=0}^q j \binom{j}{k-1}$.

From Lemma 3 and the above evaluations, we have $Q_k(m) = mQ_{k-1}(m) - P_{k-1}(m)$ for all m , since this formula is independent of d .

To finish the proof of the theorem it is sufficient to show that for $d < m \leq 2d$, $\sigma_m - c_m \geq 0$ or

$$\begin{aligned} mQ_{k-1}(d) - P_{k-1}(d) - mQ_{k-1}(m-d-1) \\ + P_{k-1}(m-d-1) - Q_k(d) + Q_k(m-d-1) \geq 0. \end{aligned}$$

Use the previous equation to simplify, obtaining as sufficient $(m-d)Q_{k-1}(d) - (d+1)Q_{k-1}(m-d-1) \geq 0$. Let $m_1 = m-d-1$; this becomes $(m_1+1)Q_{k-1}(d) \geq (d+1)Q_{k-1}(m_1)$ for $1 \leq m_1 \leq 2d-1$. Apply Lemma 2; this proves Theorem 1.

We now prove Theorem B of the Introduction, or

THEOREM 2. $\gamma(a, f) \leq \|Df(z)^{-1}\| \|f\| \varphi'_d(\|z\|)^2 / \varphi_d(\|z\|)$.

PROOF. First note $D^k f(z) = \sum_{i=0}^k i(i-1) \cdots (i-k+1) a_i z^{i-k}$ where $a_i z^{i-k}$ is the k -linear form defined by putting z into $(i-k)$ places of the i linear form a_i . From this picture it follows that

$$\left\| \frac{D^k f(z)}{k!} \right\| \leq \|f\| \frac{\varphi_d^{(k)}(\|z\|)}{k!}.$$

Since $\|Df(z)^{-1}\| \|Df(z)\| \geq \|Df(z)^{-1} Df(z)\| = 1$, using $k = 1$ we obtain $\|Df(z)^{-1}\| \|f\| \varphi'_d(\|z\|) \geq 1$. From Theorem 1, it follows for $k > 1$ that

$$\frac{\varphi_d^{(k)}(r)}{k!} \leq \frac{\varphi'_d(r)^k}{\varphi_d(r)^{k-1}}, \quad \left\| \frac{D^k f(z)}{k!} \right\| \leq \|f\| \frac{\varphi'_d(\|z\|)^k}{\varphi_d(\|z\|)^{k-1}}.$$

Now we use the above estimate to prove Theorem 2 as follows:

$$\begin{aligned} \gamma(z, f) &\leq \sup_{k \geq 1} \left\| Df(z)^{-1} \frac{D^k f(z)}{k!} \right\|^{1/(k-1)} \\ &\leq \sup_k \left(\|Df(z)^{-1}\| \left\| \frac{D^k f(z)}{k!} \right\| \right)^{1/(k-1)} \\ &\leq \sup_{k \geq 1} \left(\|Df(z)^{-1}\| \|f\| \frac{\varphi'_d(\|z\|)^k}{\varphi_d(\|z\|)^{k-1}} \right)^{1/(k-1)} \\ &\leq \frac{\varphi'_d(\|z\|)}{\varphi_d(\|z\|)} \sup_{k \geq 1} (\|Df(z)^{-1}\| \|f\| \varphi'_d(\|z\|))^{1/(k-1)}. \end{aligned}$$

The sup is over the $(k-1)$ roots of a number which is independent of k and not less than 1. Thus the sup is achieved at $k = 2$. This yields Theorem 2.

Remark. There is a kind of “ L^2 version” of Theorem 1, which I have not been able to resolve.

Problem. Is $\alpha(r, \psi_d) \leq 1$ for all $r \geq 0$ where $\psi_d(r) = (\sum_{i=0}^d r^{2i})^{1/2}$?

2. The goal of this section is to prove the tractability of the Random Algorithm by giving a probability estimate of the function σ_f (the theorem is stated in the Introduction).

Let B_f be the area of the set of $z \in D_3$ such that $\alpha(z, f) < \alpha_0$. The following is similar to Proposition 3, §2 of Shub-Smale [44] (and very simple).

PROPOSITION 1. $\sigma_f = 9\pi/B_f$. (Note area of $D_3 = 9\pi$.)

We next prove a proposition which shows that our criterion for approximate zeros of the second kind (Theorem C) also works for approximate zeros (of the first kind).

PROPOSITION 2. *There is a constant $c > 0$ with the following property. Let $f: \mathbf{E} \rightarrow \mathbf{F}$ be an analytic map from one Banach space to another and ζ an element of \mathbf{E} with $f(\zeta) = 0$. If $z \in \mathbf{E}$ satisfies $\|z - \zeta\| < c/\gamma(\zeta, f)$ then $\alpha(z, f) < \alpha_0$ and z is an approximate zero.*

PROOF. Suppose $c < \frac{1}{2}(1 - \sqrt{2}/2)$ and z satisfies the hypothesis of the proposition. Then by Lemma 2c, §3 of [P.E.], $\gamma(z) < c_1 \gamma(\zeta)$ for a suitable constant c_1 . Next use Proposition 3, §4 of [P.E.], to obtain that $\beta(z) < c_2 \|z - \zeta\|$ for a suitable constant c_2 . Choose

$$c < \min \left(\frac{1}{2} \left(1 - \frac{\sqrt{2}}{2} \right), \frac{\alpha_0}{c_1 c_2} \right).$$

So $\alpha(z, f) = \beta(z) \gamma(z) < c_1 c_2 \|z - \zeta\| \gamma(\zeta) < \alpha_0$. Q.E.D.

For each polynomial f , define

$$\rho_f = \min_{\substack{\theta \\ f'(\theta)=0}} |f(\theta)|.$$

Let $\tau_f = \min(1, \rho_f)$. From Smale [50, p. 29],

LEMMA 1. $\text{meas}\{f \in P_d(1) | \tau_f < \rho\} < d\rho^2$.

The following was proved by Shub-Smale [43, p. 124]. The new proof here does not use the theory of Schlicht functions, in contrast.

PROPOSITION 3. *There is a constant $K > 0$ such that for all $f \in P_d(1)$, $B_f > K\tau_f^2/d^4$.*

We use several lemmas on our way to proving Proposition 3. However, first observe the well-known fact that if $f \in P_d(1)$ and $f(\zeta) = 0$, then $|\zeta| < 2$. This eventually is the reason for our choice of the 3 in the theorem we are proving in this section.

LEMMA 2. *If $f \in P_d(1)$,*

$$B_f > c \sum_{\substack{\zeta \\ f(\zeta)=0}} \left(\frac{|f'(\zeta)|}{\varphi'_d(|\zeta|)} \right)^2 \left(\frac{\varphi_d(|\zeta|)}{\varphi'_d(|\zeta|)} \right)^2.$$

PROOF. It follows from Proposition 2 that

$$B_f > c \sum_{\substack{\zeta \\ f(\zeta)=0}} \frac{1}{\gamma(\zeta, f)^2}$$

for some constant $c > 0$ (compare Lemma 1 of §3). Now apply Theorem B. Since $f \in P_d(1)$, $\|f\| = 1$. This yields Lemma 2.

LEMMA 3. $\varphi'_d(r)/\varphi_d(r) \leq d$, all $r \geq 0$.

PROOF. $\sum_0^d i r^{i-1} \leq d \sum_0^d r^{i-1} \leq d \sum_0^d r^i$.

The following is just an application of the inequality on arithmetic and geometric means.

LEMMA 4.

$$\frac{1}{d} \sum_{\substack{\zeta \\ f(\zeta)=0}} \left| \frac{f'(\zeta)}{\varphi'_d(|\zeta|)} \right|^2 \geq \prod_{\substack{\zeta \\ f(\zeta)=0}} \left(\frac{|f'(\zeta)|}{\varphi'_d(|\zeta|)} \right)^{2/d}.$$

LEMMA 5. *There is a positive constant c such that*

$$\prod_{\substack{\zeta \\ f(\zeta)=0}} \varphi'_d(|\zeta|)^{1/d} \leq c d^{5/2}, \quad \text{all } f \in P_d(1).$$

PROOF. We use a theorem of Specht (see Marden [32, p. 129]) which easily implies the following assertion. If $f \in P_d(1)$ has roots ζ_1, \dots, ζ_d , then for any subset i_1, \dots, i_k of $\{1, \dots, d\}$, $|\zeta_{i_1} \zeta_{i_2} \cdots \zeta_{i_k}| \leq \sqrt{d+1}$.

Then

$$\begin{aligned}
 \prod_{\substack{\zeta \\ f(\zeta)=0}} \varphi'_d(|\zeta|) &= \prod_{i=1}^d \sum_{j=1}^d j |\zeta_i|^{j-1} \\
 &= \sum_{\substack{(i_1, \dots, i_d) \\ 1 \leq i_k \leq d}} i_1 |\zeta_1|^{i_1-1} i_2 |\zeta_2|^{i_2-1} \dots i_d |\zeta_d|^{i_d-1} \\
 &\leq (\sqrt{d+1})^d \sum i_1 i_2 \dots i_d = (\sqrt{d+1})^d \sum_1^d i_1 \sum_1^d i_2 \dots \sum_1^d i_d \\
 &\leq (\sqrt{d+1})^d \left(\frac{d(d+1)}{2} \right)^d.
 \end{aligned}$$

Then take the d th root to obtain

$$\prod_{\substack{\zeta \\ f(\zeta)=0}} (\varphi'_d(|\zeta|))^{1/d} \leq cd^{5/2}. \quad \text{Q.E.D.}$$

LEMMA 6. Let $f \in P_d(1)$ and D_f be the discriminant of f (Lang [29]). Then $|D_f|^{1/d} \geq d\tau_f$.

For this see Shub-Smale [44, Proposition 7 of §2] for the argument.

From Lemmas 2 to 6 we prove Proposition 3. By Lemmas 2 and 3

$$B_f > c \frac{1}{d^2} \sum_{\substack{\zeta \\ f(\zeta)=0}} \left(\frac{|f'(\zeta)|}{\varphi'_d(|\zeta|)} \right)^2.$$

Then by Lemmas 4 and 5 we obtain

$$B_f > c_1 \frac{1}{d} \frac{1}{d^5} \prod_{\zeta} |f'(\zeta)|^{2/d}.$$

Recalling $D_f = \prod_{\zeta} (f'(\zeta))$ (Lang [29]), Lemma 6 applies to yield $B_f > c_2 \tau_f^2 / d^4$, proving Proposition 3.

Now the theorem is proved as follows. Define $G_d: Z^+ \rightarrow [0, 1]$ by

$$G_d(\sigma) = \text{meas}\{f \in P_d(1) | \sigma_f > \sigma\}.$$

The theorem may be restated: $G_d(\sigma) < cd^5/\sigma$. But using Propositions 1 and 3

$$\begin{aligned}
 G_d(\sigma) &= \text{meas}\{f \in P_d(1) | 9\pi/B_f > \sigma\}, \\
 G_d(\sigma) &\leq \text{meas}\{f \in P_d(1) | cd^4/\tau_f^2 > \sigma\}
 \end{aligned}$$

or

$$G_d(\sigma) \leq \text{meas}\{f \in P_d(1) | \tau_f < (cd^4/\sigma)^{1/2}\}.$$

Now use Lemma 1, so $G_d(\sigma) \leq cd^5/\sigma$. Q.E.D.

3. The goal of this section is to prove the theorem on the Average Area of Approximate Zeros.

The quantity $\text{average}_{f \in P_d(1)} A_f$ by definition is equal to

$$\left(\frac{1}{\pi}\right)^d \int_{|a_{d-1}| \leq 1} \cdots \int_{|a_0| \leq 1} A_f da_0 \cdots da_d.$$

Our first task will be to make the estimate

$$\int_{|a_0| \leq 1} A_f da_0 \geq K \int_{\substack{|\zeta| < 1/2 \\ \zeta \in \mathbb{C}}} |f'(\zeta)|^4 d\zeta. \quad (\text{I})$$

Then, by (I), for the proof of the theorem, it is sufficient to show that

$$\int_{|a_{d-1}| \leq 1} \cdots \int_{|a_1| \leq 1} \int_{|\zeta| < 1/2} |f'(\zeta)|^4 d\zeta > L(\pi)^d. \quad (\text{II})$$

Of course K and L are positive constants, independent of d . Some estimates of these constants can be given from the proof.

First we deal with (I). Let $\gamma_\zeta = \max(1, \gamma(\zeta, f))$ where $\gamma(\zeta, f)$ is as in the Introduction.

LEMMA 1.

$$A_f > \left(\frac{3 - \sqrt{7}}{2}\right)^2 \pi \sum_{\substack{\zeta \\ f(\zeta)=0 \\ |\zeta| < 1/2}} \gamma_\zeta^{-2}.$$

PROOF OF LEMMA 1. Let $\hat{\Omega}_f(\zeta)$ be the set of approximate zeros z of the second kind corresponding to an actual zero such that $|z - \zeta| < \frac{1}{2}\zeta$. The $\hat{\Omega}_f(\zeta)$ are disjoint and

$$\Omega_f \supset \bigcup_{|\zeta| < 1/2} \hat{\Omega}_f(\zeta).$$

From Theorem C, it follows that $z \in \hat{\Omega}_f(\zeta)$ if $\|z - \zeta\| < (3 - \sqrt{7})/2\gamma_\zeta$. Therefore the area of $\hat{\Omega}_f(\zeta)$ is greater than or equal to $\pi((3 - \sqrt{7})/2\gamma_\zeta)^2$. Putting these facts together yields Lemma 1.

Now we prove the following change of variable formula.

LEMMA 2. Let $f_0 = \sum_{i=1}^d a_i z^i$, with $|a_i| \leq 1$, so that $f_0(0) = 0$, and consider $f = f_0 + a_0$, a_0 varying in \mathbb{C} with $|a_0| \leq 1$. Then

$$\int_{|a_0| \leq 1} \sum_{\substack{\zeta \\ f(\zeta)=0 \\ |\zeta| < 1/2}} \gamma_\zeta^{-2} da_0 = \int_{\substack{\zeta \in f_0^{-1}(D_1) \\ |\zeta| < 1/2}} \gamma_\zeta^{-2} |f'_0(\zeta)|^2 d\zeta.$$

PROOF OF LEMMA 2. Note that $|f_0(\zeta)| < 1$ if $|\zeta| < \frac{1}{2}$. Let $V = f_0(D_{1/2})$ and let $g_i: V \rightarrow D_{1/2}$, $i = 1, \dots, d$, be branches of f_0^{-1} restricted to $D_{1/2}$. If $\mathcal{U}_i = g_i(V)$, then the \mathcal{U}_i are disjoint measurable subsets of $D_{1/2}$ whose union is $D_{1/2}$, and $f_0 g_i(w) = w$ for all w in V .

Note that $\zeta = g_i(-a_0)$ if and only if $f_0(\zeta) = -a_0$, or equivalently, $f(\zeta) = 0$ when $f = f_0 + a_0$. Then

$$\begin{aligned} \int_{\zeta \in D_{1/2}} \gamma_\zeta^{-2} |f'_0(\zeta)|^2 d\zeta &= \sum_{i=1}^d \int_{\zeta \in \mathcal{U}_i} \gamma_\zeta^{-2} |f'_0(\zeta)|^2 d\zeta \\ &= \sum_{i=1}^d \int_{a_0 \in V} (\gamma(g_i(-a_0)))^2 da_0 \quad (\text{change of variable}) \\ &= \int_{a_0 \in V} \sum_{f(\zeta)=0} \gamma(\zeta)^{-2} da_0 \\ &= \int_{|a_0| \leq 1} \sum_{\substack{f(\zeta)=0 \\ |\zeta| < 1/2}} \gamma(\zeta)^{-2} da_0. \end{aligned}$$

LEMMA 3. *There is a constant $K_1 > 0$ such that $\gamma_\zeta^{-2} \geq K_1 |f'(\zeta)|^2$ if $|\zeta| < \frac{1}{2}$.*

PROOF. Suppose $\gamma_\zeta = \gamma(\zeta, f)$. Use Theorem B so

$$\gamma_\zeta \leq \frac{\varphi'_d(|\zeta|)^2}{\varphi_d(|\zeta|)} \|f\| \frac{1}{|f'(\zeta)|}.$$

Therefore

$$\gamma_\zeta^2 \leq \frac{\varphi'_d(1/2)^4}{\varphi_d(1/2)^2} \frac{1}{|f'(\zeta)|^2}$$

for $f \in P_d$ and $|\zeta| \leq \frac{1}{2}$.

The case of $\max(1, \gamma(\zeta, f)) = 1$ is easily checked directly.

This yields Lemma 3 with K_1 easily estimated explicitly. Then Lemmas 1, 2, and 3 yield (I); just note $f' = f'_0$.

Toward the proof of (II), we have

LEMMA 4. *For any complex polynomial g ,*

$$\int_{\substack{|z| < R \\ z \in \mathbb{C}}} g(z) \overline{g(z)} dz \geq |g(0)|^2 \pi R^2.$$

PROOF. Use polar coordinates. The odd powers of 2 give a zero contribution and the even powers a positive contribution. The lemma follows easily.

Now apply Lemma 4 to obtain (II). Here $g(\zeta) = f'(\zeta)^2$ and $g(0) = (a_1)^2$. This gives the first step into the integration of (II). Next one integrates over $|a_1| \leq 1$ using polar coordinates. The remaining integrations are trivial. This proves (II) and hence the theorem.

Another perhaps even simpler proof of the analogous theorem, using approximate zeros (of the first kind), goes by Theorem A, Lemmas 1 and 3 above, and

LEMMA 5. *There exist universal positive constants δ and ε such that if $f(z) = \sum_{i=0}^d a_i z^i$, $|a_i| \leq 1$, $|a_0|/|a_1|^2 < \varepsilon$, and $|z| < \delta$, then $\alpha(z, f) < \alpha_0$ and so z is an approximate zero (of the first kind).*

4. Here we give the proof of the theorem on the speed of a Global Newton Method.

LEMMA 1. *There is a positive constant K with this property. Given any (z_0, f) with $\alpha(z_0, f) < \alpha_0$, let $z_l = N_f^l(z_0)$, $l = 1, 2, \dots$, and $\zeta = \lim_{l \rightarrow \infty} z_l$. Then $\gamma(\zeta, f) \leq K\gamma(z_0, f)$.*

PROOF OF LEMMA 1. Let $\alpha_l = \alpha(z_l)$, $\psi_l = \psi(\alpha_l)$, $l = 1, 2, \dots$, where $\psi(r) = 2r^2 - 4r + 1$. Then α_l decreases as l increases and $\alpha_l \leq (\frac{1}{2})^{2^l - 1} \alpha(z_0)$ (see [P.E.], §4, Proposition 1 and §3, Proposition 2). Therefore $1/\psi_l$ decreases to 1 from above (and very fast) with l (see the beginning of §4 in [P.E.]).

Next define

$$K_0 = \frac{1}{(1 - \alpha(z_0))\psi(\alpha(z_0))} \quad \text{and} \quad K_l = \frac{1}{(1 - \alpha_l)\psi_l}, \quad l = 1, 2, \dots$$

By Lemma 2c, §3 of [P.E.], $\gamma(z_l) \leq K_{l-1}\gamma(z_{l-1})$ so $\gamma(z_l) \leq \gamma(z_0) \prod_{j=0}^{l-1} K_j$. Then it is sufficient to show that $\prod_{j=0}^{l-1} K_j$ increases to a finite constant K as $l \rightarrow \infty$.

This follows from

$$\log \prod_0^{l-1} K_j = \prod_0^{l-1} \log K_j = \sum_0^{l-1} -\log \psi_l - \sum_0^{l-1} \log(1 - \alpha_l)$$

and our estimate on α_l above. Q.E.D.

Let $K_2 < 1\frac{3}{4}$ be as in Proposition 1 of §2 of [P.E.], and define $L_0 = KK_2\alpha_0$ where K is as in Lemma 1.

LEMMA 2. *As in Lemma 1, let $\alpha(z_0, f) < \alpha_0$, $z_l = N_f^l(z_0)$, and $\zeta = \lim z_l$. Then $\|z_l - \zeta\|\gamma(\zeta) < (\frac{1}{2})^{2^l - 1} L_0$.*

PROOF OF LEMMA 2. By Proposition 1, §3 of [P.E.],

$$\|z_l - \zeta\| < K_2 \left(\frac{1}{2}\right)^{2^l - 1} \|z_1 - z_0\|.$$

Then by Lemma 1

$$\|z_l - \zeta\|\gamma(\zeta) < K_2 \left(\frac{1}{2}\right)^{2^l - 1} \|z_1 - z_0\|\gamma(z_0)K.$$

Lemma 2 follows from the definition of L_0 and the fact that $\|z_1 - z_0\|\gamma(z_0) = \alpha(z_0) < \alpha_0$.

Next we will prove the theorem. It is sufficient to prove that $\alpha(z_i, f - w_{i+1}) < \alpha_0$ for $i = 0, 1, \dots, n-1$, and for this we proceed inductively. First observe

$$\begin{aligned} \alpha(z_i, f - w_{i+1}) &\leq \gamma(z_i) \|Df(z_i)^{-1}(f(z_i) - w_i)\| \\ &\quad + \gamma(z_i) \|Df(z_i)^{-1}(w_i - w_{i+1})\|. \end{aligned}$$

It is sufficient to show that (a) and (b) are true for a suitable fixed l , where

- (a) $\gamma(z_i) \|Df(z_i)^{-1}(f(z_i) - w_i)\| \leq \alpha_0 \left(\frac{1}{2}\right)^{2^l - 1}$,
- (b) $\gamma(z_i) \|Df(z_i)^{-1}(w_i - w_{i+1})\| \leq \alpha_0 (1 - \left(\frac{1}{2}\right)^{2^l - 1})$.

For $i = 0$, (a) is true since $f(z_0) = w_0$. For $i > 0$, use [P.E.], Proposition 1b of §4, with $a = \frac{1}{2}$ to obtain $\alpha(z_i, f - w_i) \leq (\frac{1}{2})^{2^i-1} \alpha(z_{i-1}, f - w_i)$ which, by the induction hypothesis, is less than $(\frac{1}{2})^{2^i-1} \alpha_0$. That proves (a).

For the proof of (b), define $\zeta_i = \lim_{k \rightarrow \infty} N_{f-w_i}^k(z_{i-1})$ so that $f(\zeta_i) = w_i$, and let $\tau = \|z_i - \zeta_i\| \gamma(\zeta_i)$. By Lemma 2, $\tau \leq (\frac{1}{2})^{2^i-1} L_0$, and it can be seen that for l larger than some small l_0 ($= 4$ for example), $\tau \leq 1 - \sqrt{2}/2$.

By Lemma 2c, §3 of [P.E.],

$$\gamma(z_i) \leq \gamma(\zeta_i) \left(\frac{1}{1-\tau} \right) \frac{1}{\psi(\tau)}.$$

Also using Lemma 2b, §2 of [P.E.],

$$\begin{aligned} \|Df(z_i)^{-1}(w_i - w_{i+1})\| &\leq \|Df(z_i)^{-1} Df(\zeta_i)\| \|Df(\zeta_i)^{-1}(w_i - w_{i+1})\| \\ &\leq \frac{(1-\tau)^2}{\psi(\tau)} \beta(\zeta_i, f) \frac{\|w_i - w_{i+1}\|}{\|w_i\|}. \end{aligned}$$

Therefore, for (b) it is sufficient to show

$$\gamma(\zeta_i) \frac{1-\tau}{\psi(\tau)^2} \beta(\zeta_i, f) \left(\frac{w_i - w_{i+1}}{w_i} \right) < \alpha_0 \left(1 - \left(\frac{1}{2} \right)^{2^i-1} \right)$$

or

$$\frac{\alpha(\zeta_i, f)}{\|w_i\|} \frac{(1-\tau)}{\psi(\tau)^2} \Delta \|f(z_0)\| < \alpha_0 \left(1 - \left(\frac{1}{2} \right)^{2^i-1} \right)$$

or

$$M \Delta \|f(z_0)\| \frac{(1-\tau)}{\psi(\tau)^2} < \alpha_0 \left(1 - \left(\frac{1}{2} \right)^{2^i-1} \right)$$

or, for some $c > 0, l$,

$$\frac{1}{c} \left[\frac{1-\tau}{\psi(\tau)^2} \right]_{\tau=(1/2)^{2^l-1} L_0} \leq \alpha_0 \left(1 - \left(\frac{1}{2} \right)^{2^l-1} \right).$$

The last equation is clear and defines a choice of constants c and l of the theorem. Q.E.D.

5. We prove here the theorem on the regularization of the LPP. Then we add some discussion related to Newton method algorithms for this problem.

Toward the proof of the theorem we have

LEMMA 1. *The inverse of the derivative $D\Phi_a(x)$: $\mathbf{R}^N \rightarrow \mathbf{R}^N$ exists and has a bound on its norm independent of M ,*

$$\|D\Phi_a(x)^{-1}\| \leq 4 \max_i \left(\frac{x_i^2 + a^2}{a^2} \right).$$

REMARK. This already shows the continuity of the solution of our perturbed LPP in (A, b, c) .

PROOF OF LEMMA 1. First note that M is antisymmetric so that $M^T x = -Mx$, M^T the transpose of M . This is clear from the structure of M as

$M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}$. Thus for any $v \in \mathbf{R}^N$, $\langle Mv, v \rangle = 0$ where $\langle \cdot, \cdot \rangle$ is the usual inner product on \mathbf{R}^n .

This fact is crucial for all the subsequent analysis (although the milder condition $\langle Mv, v \rangle \geq 0$ is sufficient).

Fix the following notation for the rest of this section.

$$\Delta = \text{diag matrix} \left(\frac{x_i}{(x_i^2 + a^2)^{1/2}} \right)_{i=1}^N$$

so $2D\Phi_a(x) = I + \Delta + M(I - \Delta)$.

The above observation yields $\langle 2D\Phi_a(x)(u), (I - \Delta)u \rangle = \langle (I + \Delta)u, (I - \Delta)u \rangle$. Applying the Schwartz inequality to the left side gives us

$$|\langle (I + \Delta)u, (I - \Delta)u \rangle| \leq 2\|D\Phi_a(x)(u)\| \|(I - \Delta)u\|.$$

The left side of the last estimate equals $\sum u_i^2(a^2/(x_i^2 + a^2))$; moreover $\|(I - \Delta)u\| \leq 2\|u\|$. Thus

$$\min_i \left(\frac{a^2}{x_i^2 + a^2} \right) \|u\| \leq 4\|D\Phi_a(x)(u)\|.$$

The last estimate is sufficient to yield Lemma 1.

LEMMA 2. For real numbers a, x, y ,

$$\begin{aligned} & (x - y)^2 - ((x^2 + a^2)^{1/2} - (y^2 + a^2)^{1/2})^2 \\ &= 2((xy + a^2)^2 + a^2(x - y)^2)^{1/2} - 2(xy + a^2). \end{aligned}$$

The proof just goes by expanding out and comparing terms. Note that the right-hand side is always positive unless $x = y$. Thus the same is true for the left-hand side.

LEMMA 3. The map $\Phi_a: \mathbf{R}^N \rightarrow \mathbf{R}^N$ is one to one if $a > 0$.

PROOF. Let $\Lambda_x = ((x_1^2 + a^2)^{1/2}, \dots, (x_N^2 + a^2)^{1/2})$, so $2\Phi_a(\Lambda) = x + \Lambda_x + M(x - \Lambda_x)$.

We have

$$\begin{aligned} & 2\langle \Phi_a(x) - \Phi_a(y), (x - \Lambda_x) - (y - \Lambda_y) \rangle \\ &= \langle x + \Lambda_x - (y + \Lambda_y), (x - \Lambda_x) - (y - \Lambda_y) \rangle \\ &= \sum_{i=1}^N (x_i - y_i)^2 - ((x_i^2 + a^2)^{1/2} - (y_i^2 + a^2)^{1/2})^2. \end{aligned}$$

Now apply Lemma 2 to finish the proof of Lemma 3.

LEMMA 4. The image of $\Phi_a: \mathbf{R}^N \rightarrow \mathbf{R}^N$ is contained in \mathcal{U}_M .

PROOF. Since $\Phi_a(x) = \Phi_a^+(x) + M\Phi_a^-(x)$, $\Phi_a(x)$ is a positive linear combination of the coordinate vectors e_1, \dots, e_N and the columns of $-M$. Even more explicitly, note that

$$\varphi_a^+(x_i)\varphi_a^-(x_i) = -a^2/4, \quad i = 1, \dots, N.$$

If M_k is the k th column of M ,

$$\Phi_a(x) = \sum_k \left(\varphi_a^+(x_k) e_k + \frac{a^2}{4\varphi_a^+(x_k)} (-M_k) \right).$$

Now consider the usual LCP which, as always, we suppose is generated by the LPP. A result of Cottle [52] (the Theorem on p. 663) can be interpreted as

LEMMA 5. *Given M , $\Phi_M(x) = q$ has a solution x if and only if q belongs to the closure $\bar{\mathcal{U}}_M$ of \mathcal{U}_M .*

LEMMA 6. *Given M and $q \in \mathcal{U}_M$, there is an a^* such that for each $0 < a < a^*$ there is a solution x of $\Phi_a(x) = q$.*

The proof of Lemma 6 goes by using a topological degree argument and Lemma 5. Let $V = V_\delta(q)$ be the closed ball of radius δ about q in \mathbf{R}^N , with δ chosen so that $V_\delta(q) \subset \mathcal{U}_M$. Then from standard linear programming theory, the restriction $\Phi_M / : \Phi_M^{-1}(V) \rightarrow V$ is well defined, proper, and of degree 1. Therefore, since Φ_a uniformly approximates Φ_M , we obtain the conclusion of Lemma 6.

To finish the proof of the theorem, it is sufficient to show that the map $\Phi_a : \mathbf{R}^N \rightarrow \mathcal{U}_M$ has image all of \mathcal{U}_M . This argument proceeds as follows. Fix M, q and consider the solution $x(a)$ of $\Phi_a(x(a)) = q$, given by $x(a) = \Phi_a^{-1}(q)$ for small a , by Lemma 6. Thus $x(a)$ is defined for $a < a^*(q)$ where $a^*(q)$ may be supposed maximal. We will prove that $a^*(q) = \infty$.

For each q (fixed M), $x(a)$ satisfies an ordinary differential equation by differentiating $\Phi_a(x(a)) = q$ with respect to a . This equation is

$$\frac{dx}{da} = -D\Phi_a(x)^{-1} \frac{\partial \Phi_a}{\partial a}.$$

Here

$$2 \frac{\partial \Phi_a}{\partial a} = (I - M) \left(\frac{a}{(x_i^2 + a^2)^{1/2}} \right)_{i=1}^N$$

and $2D\Phi_a = (I + \Delta) + M(I - \Delta)$ as above.

Let $u = dx/da$. Then u satisfies

$$(I + \Delta)u + \frac{a}{(x_i^2 + a^2)^{1/2}} \Big|_{i=1}^N + M \left\{ (I - \Delta)u - \frac{a}{(x_i^2 + a^2)^{1/2}} \Big|_{i=1}^N \right\} = 0.$$

Take the inner product with the quantity in braces to obtain

$$\left\langle (I + \Delta)u + \frac{a}{(x_i^2 + a^2)^{1/2}} \Big|_i, (I - \Delta)u - \frac{a}{(x_i^2 + a^2)^{1/2}} \Big|_i \right\rangle = 0$$

or $\|u\|^2 = \|\Delta u + a/(x_i^2 + a^2)^{1/2}\|_i^2$. This implies

$$\sum_{i=1}^N \frac{(au_i - x_i)^2 - (x_i^2 + a^2)}{(x_i^2 + a^2)} = 0$$

or yet

$$\sum_{i=1}^N \frac{(u_i - x_i/a)^2}{1 + (x_i/a)^2} = N.$$

Thus for each i ,

$$|u_i - x_i/a| \leq \sqrt{N}(1 + (x_i/a)^2)^{1/2}$$

and so u_i satisfies an estimate of the form $|u_i| \leq K_1 + K_2|x_i/a|$. This bound implies our $a^*(q) = \infty$, the surjectivity of $\Phi_a: \mathbf{R}^N \rightarrow \mathcal{U}_M$ and hence the theorem.

Next we show how the theorems on approximate zeros can be used to initiate algorithms for solving our approximation of the LPP.

Note that

$$\Phi_a(0) = \frac{a(I - M)}{2}q_0, \quad q_0 = (1, \dots, 1).$$

Define an associated map $\hat{\Phi}_a(x) = \Phi_a(x) - \Phi_a(0)$. Then if $\hat{\Phi}_a(x(a)) = q$, as $a \rightarrow 0$, $x(a)$ tends to a solution of $\Phi_M(x) = q$ and hence the original LPP. Moreover, $x(a)$ satisfies the ordinary differential equation

$$\frac{dx}{da} = -D\hat{\Phi}_a(x)^{-1} \frac{\partial \hat{\Phi}_a}{\partial a}(x).$$

To find an initial value for this equation (to let a decrease to 0) it is sufficient to solve $\hat{\Phi}_a(x) = q$ for some value of a . This motivates a

PROPOSITION. *If $a \geq 8||q||/\alpha_0$, then $0 \in \mathbf{R}^N$ is an approximate zero of $\hat{\Phi}_a - q$.*

Thus for $a = 8||q||/\alpha_0$, one can solve $\hat{\Phi}_a(x) = q$ with great precision in a few (e.g., 5) Newton iterations starting at 0. Here recall $\alpha_0 > \frac{1}{8}$.

We use Theorem A and show that $\alpha(0, \hat{\Phi}_a - q) < \alpha_0$.

LEMMA 7. *Let*

$$\varphi(x) = \varphi_a^\pm(x) = \frac{x \pm (x^2 + a^2)^{1/2}}{2}.$$

Then $\alpha(x, \varphi) \leq 4$, and for all $k > 1$

$$\left(\frac{2|\varphi^{(k)}(0)|}{k!} \right)^{1/(k-1)} \leq \frac{4}{a}.$$

PROOF. Our proof uses Schlicht function theory (in particular, Loewner's work on the Bieberbach Conjecture)! An elementary estimate of the higher derivatives can be done, but is not trivial. The proof assumes the plus sign, but with φ_a^- the proof is similar.

First note that the inverse of $\varphi: \mathbf{R} \rightarrow \mathbf{R}^+$ is given by $g(y) = y - a^2/4y$. Thus the radius of convergence of the inverse about y is y or, in terms of x , just $\varphi(x)$.

A formula of Smale [50], Shub-Smale [43], or Smale [51, p. 105] (proved using work of Loewner) may be stated in the form $\alpha(x, \varphi) \leq 4|\varphi(x)|/r(\varphi_x^{-1})$ for quite general φ . Here $r(\varphi_x^{-1})$ is the radius of convergence of φ^{-1} about $\varphi(x)$.

Thus $\alpha(x, \varphi) \leq 4$. Moreover, $\varphi(0) = \frac{a}{2}$ and $\varphi'(0) = \frac{1}{2}$ so the last part of Lemma 7 follows. Elise Cawley recently showed that by a direct analysis of the k th derivative one obtains $\frac{1}{4a}$, not $\frac{4}{a}$. This improves the 8 in the Proposition to $\frac{1}{2}$.

LEMMA 8. *If M is antisymmetric, then $\|(I + M)^{-1}(I - M)\| = 1$.*

PROOF. Consider $(I + M)^{-1}(I - M)u = v$. Then $(u - v) - M(u + v) = 0$ and $\langle u - v, u + v \rangle = 0$ so $\|u\| = \|v\|$. Q.E.D.

LEMMA 9. *Let $G: \mathbf{R}^N \rightarrow \mathbf{R}^N$ be an analytic map of the form $G(x) = (g_1(x_1), \dots, g_n(x_n))$. Then*

$$\|D^k G(x)\| = \frac{\sum_1^N \lambda_i^2}{(\sum_1^N \lambda_i^{2/k})^{k/2}}$$

where $\lambda_i = g_i^{(k)}(x_i)$.

PROOF. For $v \in \mathbf{R}^N$,

$$D^k G(x)(v^k) = (\lambda_1 v_1^k, \dots, \lambda_N v_N^k), \quad \|D^k G(x)(v^k)\| = \left(\sum (\lambda_i v_i^k)^2 \right)^{1/2}$$

and the maximum over $\|v\| = 1$ is at $v_i^k = K \lambda_i$ for some $K > 0$. So

$$\|D^k G(x)(v^k)\| = K \left(\sum \lambda_i^2 \right)^{1/2}, \quad K = 1 / \left(\sum \lambda_i^{2/k} \right)^{k/2}. \quad \text{Q.E.D.}$$

Now we can easily make the estimate on $\alpha(0, \hat{\Phi}_a - q)$. We have $\hat{\Phi}_a(0) = 0$, $D\hat{\Phi}_a(0) = (I + M)/2$, $\|D\hat{\Phi}_a(0)^{-1}\| \leq 2$, so that $\beta(0, \hat{\Phi}_a - q) \leq 2\|q\|$. Moreover,

$$\begin{aligned} \gamma(0, \hat{\Phi}_a - q) &\leq \sup_{k \geq 1} \left\| D\hat{\Phi}_a^{-1}(0) \frac{D^k \hat{\Phi}_a(0)}{k!} \right\|^{1/(k-1)} \\ &\leq \left\| (I + M)^{-1}(I - M) \frac{D^{k-1} \Delta}{k!} \right\|^{1/(k-1)} \\ &\leq \left| \frac{\varphi_a^{(k)}(0)}{k!} \right|^{1/(k-1)} \leq \frac{4}{a}. \end{aligned}$$

We have used Lemmas 7, 8, and 9. Thus $\alpha = \beta\gamma \leq 8\|q\|/a < \alpha_0$, proving the proposition.

6. In Smale [51], one dozen problems in the algorithms of analysis were explicitly listed. Since that paper was written, solutions or progress for half of them have resulted. In this section we report on these results.

Problem 1 deals with extending work of Smale [50], Shub-Smale [43, 44], and on zero finding of polynomials. This work proves probability bounds for algorithms related to Newton's method, and eventually, that the number of iterations is proportional to the degree of the polynomial. This work used the theory of Schlicht functions (related to the Bieberbach Conjecture) and did not extend to finding zeros of polynomial systems $f: \mathbf{C}^n \rightarrow \mathbf{C}^n$ where $n > 1$. Problem 1

asked for such results. Renegar [40] has now found polynomial bounds, statistically, for each n . This important work thus gives some answer to our problem. However, deep questions, including those on the conceptual level, remain. The bounds of Renegar, while polynomial in the degrees, are crude. For one variable, for example, the number of iterations is on the order of d^{26} (in contrast to d of Shub-Smale [44]). Renegar's results are one of the inspirations for this paper, and I hope that the theorems here will make some contribution to the problem of complexity for polynomial systems.

Problem 6 deals with the question of understanding systematically the set of polynomials of one variable, where Newton's method can cycle on open sets. Cayley showed that for quadratic polynomials, this could not happen generically. Janet Head [18] of Cornell has given a good analysis of the situation of cubics and has shown that cycling in that case is a rather rare occurrence. Her work does not include f of degree higher than 3.

Problem 7 raises the problem as to how often (simple) Newton's method converges for polynomials of one variable. Let A_f be the normalized area of the set of points of D_2 which converges to a zero of f under Newton's method. Let

$$A_d = \min_{f \in P_d(1)} A_f.$$

Joel Friedman [12] at the University of California at Berkeley proved my conjecture that $A_d > 0$, all d , and gave some estimates on A_d as a function of d , dealing well with Problem 7A. For Problem 7B, see Problem 9 below.

Problem 8. In it, a graph Γ_f is assigned to each polynomial f which reflects the topology of f as a map, especially related to Newton's Method. The problem itself asks what graphs can occur. Work dealing with this had been already initiated by Jongen-Jonker-Twilt [22]. See their two articles for references and an account of their work. Moreover, Shub-Tischler-Williams [46] have given a good study, and one could say that the problem is essentially solved.

Problem 9, on the average area of the set of approximate zeros being bounded below, is answered completely by the theorem proved in §3 of the present paper (incidentally, answering Problem 7B as well).

Problem 10 conjectured that there exists no purely iterative generally convergent algorithms for polynomial zero solving over \mathbf{C} . Curt McMullen [33] solved the problem completely. More precisely he showed this

THEOREM. *Let $d > 3$ and $T: P_d \times S \rightarrow S$ be any map rational over \mathbf{C} in f and z . Then there is no open set $U \subset P_d \times S$ of full measure with this property: If $(f, z) \in U$, then $T_f^k(z) = z_k$ converges to a root of f as $k \rightarrow \infty$.*

Here P_d is the space of all polynomial of degree $\leq d$ and S is the Riemann sphere. That T is rational over \mathbf{C} means that T can be formed from the complex rational operations $(+, -, \times, \div)$ in the coefficients of f and z . For $d = 2$, Newton's method or $T_f(z) = z - f(z)/f'(z)$ is such an algorithm as proved by Cayley.

For $d = 3$, McMullen [33] found a new algorithm which is purely iterative and generally convergent.

In McMullen [34, 35] he deepens his investigation. In Shub-Smale [45], Shub and I showed that if one adds the operation complex conjugation, then one can construct purely iterative generally convergent algorithms. Moreover, our result holds for polynomial maps of several variables, $f: \mathbb{C}^n \rightarrow \mathbb{C}^n$.

REFERENCES

1. J. Abadie and G. Guerrero, *Méthode du GRG, méthode de Newton globale et application à la programmation mathématique*, RAIRO Rech. Opér. **18** (1984), 319–351.
2. M. Berger, *Non-linearity and functional analysis*, Academic Press, New York, 1974.
3. L. Blum, *Towards an asymptotic analysis of Karmakar's algorithm*, Inform. Process. Lett. **23** (1986), 189–194.
4. S. Chow, J. Mallet-Paret, and J. Yorke, *Finding zeros of maps: homotopy methods that are constructive with probability one*, Math. Comp. **32** (1978), 887–899.
5. R. Cottle, *Note on a fundamental theorem in quadratic programming*, J. SIAM **12** (1964), 663–665.
6. R. Cottle and G. Dantzig, *Complementary pivot theory of math. programming in math. of the decision sciences*, G. Dantzig and A. Veisott Jr., eds., Amer. Math. Soc., Providence, R.I., 1968, pp. 115–136.
7. J. Curry, *On zero finding methods of higher order from data at one point*, MSRI Preprint, 1986.
8. G. Dantzig, *Linear programming and extensions*, Princeton Univ. Press, Princeton, N.J., 1963.
9. B. Dejon and P. Henrici, *Constructive aspects of the fundamental theorem of algebra*, Wiley, New York, 1969.
10. J. Dieudonné, *Foundations of modern analysis*, Academic Press, New York, 1960.
11. C. Eaves and H. Scarf, *The solution of systems of piecewise linear equations*, Math. Oper. Res. **1** (1976), 1–27.
12. J. Friedman, *Rough draft of a theorem on Newton's method*, Univ. of California Press, Berkeley, Calif., 1985.
13. F. Gao, *Nonasymptotic error of numerical integration—an average analysis*, Univ. of California Press, Berkeley, Calif., 1986 (to appear).
14. —, *Probabilistic analysis of automatic integration* (to appear).
15. C. Garcia and F. J. Gould, *Relations between several path following algorithms and local and global Newton methods*, SIAM Rev. **22** (1980), 263–274.
16. G. Ghellinck, and M. P. Vial, *A polynomial Newton method for linear programming*, CORE, Louvain, Belgium, 1986.
17. P. Gill, W. Murray, M. Saunders, J. Tomlin, and M. Wright, *On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method*, Tech. Report SOL 85-11, Dept. of Op. Res., Stanford Univ., Stanford, Calif., 1985.
18. J. Head (to appear).
19. P. Henrici, *Applied and computational complex analysis*, Wiley, New York, 1977.
20. M. Hirsch and S. Smale, *On algorithms for solving $f(x) = 0$* , Comm. Pure Appl. Math. **32** (1979), 281–312.
21. H. Th. Jongen, P. Jonker, and F. Twilt, *The continuous desingularized Newton's method for meromorphic functions*, Memorandum No. 501, Dept. of Appl. Math., Twente Univ. of Technology, 1985.
22. —, *Non-linear optimization theory in \mathbb{R}^n from a global point of view*. VIII, Memorandum No. 559, Dept. of Appl. Math., Twente Univ. of Technology, 1986.
23. L. Kantorovich and G. Akilov, *Functional analysis in normed spaces*, Macmillan, New York, 1964.

24. N. Karmarkar, *A new polynomial-time algorithm for linear programming*, *Combinatorica* **4** (1984), 373–395.
25. H. Keller, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, *Application of Bifurcation Theory*, Academic Press, New York, 1977.
26. —, *Global homotopic and Newton methods*, *Recent Advances in Numerical Analysis*, Academic Press, New York, 1978.
27. M. Kim, Ph.D. Thesis, Graduate School, CUNY, 1985 (to appear).
28. H. Kung, *The complexity of obtaining starting points for solving operator equations by Newton's method*, *Analytic Computational Complexity* (J. Traut, ed.), Academic Press, New York, 1976.
29. S. Lang, *Algebra*, Addison-Wesley, Reading, Mass., 1963.
30. —, *Real analysis*, Addison-Wesley, Reading, Mass., 1983.
31. O. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, *SIAM J. Appl. Math.* **31** (1976), 89–92.
32. M. Marden, *Geometry of polynomials*, *Math. Surveys*, no. 3, Amer. Math. Soc., Providence, R.I., 1966.
33. C. McMullen, *Families of rational maps and iterative root-finding algorithms* (to appear).
34. —, *Automorphisms of rational maps I: Nielsen realization and dynamics on the ideal boundary*, MSRI, Berkeley, Calif., 1986.
35. —, *Automorphisms of rational maps II: braiding of the attractor and the failure of iterative algorithms*, MSRI, Berkeley, Calif., 1986.
36. N. Megiddo and M. Shub, *The boundary behavior of interior methods for linear programming* (to appear).
37. L. Nazareth, *Homotopy techniques in linear programming*, *Algorithmica* (to appear).
38. A. Ostrowski, *Solutions of equations in Euclidean and Banach spaces*, Academic Press, New York, 1973.
39. V. Pan, *Algebraic complexity of computing polynomial zeros*, Tech. Report 85-27, SUNY, Albany, N.Y., 1985.
40. J. Renegar, *On the efficiency of Newton's method in approximating all zeros of a system of complex polynomials*, *Math. Oper. Res.* (to appear).
41. —, *A polynomial-time algorithm based on Newton's method for linear programming*, MSRI, Berkeley, Calif., 1986.
42. H. Royden (to appear).
43. M. Shub and S. Smale, *Computational complexity: on the geometry of polynomials and a theory of cost. I*, *Ann. Sci. École Norm. Sup. (4)* **18** (1985), 107–142.
44. —, *Computational complexity: on the geometry of polynomials and a theory of cost. II*, *SIAM J. Comput.* **15** (1986), 145–161.
45. —, *On the existence of generally convergent algorithms*, *J. Complexity* **2** (1986), 2–11.
46. M. Shub, D. Tischler, and R. Williams, *The Newton's graph of a complex polynomial* (submitted to *SIAM J. Math. Anal.*).
47. S. Smale, *A convergent process of price adjustment and global Newton methods*, *J. Math. Econom.* **3** (1976), 107–120.
48. —, *On the average number of steps in the simplex method of linear programming*, *Math. Programming* **27** (1983), 241–262.
49. —, *The problem of the average speed of the simplex method*, *Mathematical Programming: The State of the Art* (Bonn, 1982), Springer-Verlag, Berlin and New York, 1983, pp. 530–539.
50. —, *The fundamental theorem of algebra and complexity theory*, *Bull. Amer. Math. Soc. (N.S.)* **4** (1981), 1–36.
51. —, *On the efficiency of algorithms of analysis*, *Bull. Amer. Math. Soc. (N.S.)* **13** (1985), 87–121.
52. —, *Newton's method estimates from data at one point*, *Proceedings of a Conference in Honor of Gail Young*, Laramie, Springer, New York, 1986 (to appear). (Referred to as [P.E.] in text.)

- 53. A. Wierzbicki, *Note on the equivalence of Kuhn-Tucker complementarity conditions to an equation*, J. Optim. Theory Appl. **37** (1982), 401–405.
- 54. S. Wong, *Newton's method and symbolic dynamics*, Proc. Amer. Math. Soc. **91** (1984), 245–253.
- 55. H. Woźniakowski, *A survey of information-based complexity*, J. Complexity **1** (1985), 11–44.
- 56. Paul Wright, *Statistical complexity of the power method for Markov chains*, Univ. Calif., Berkeley, 1986. (to appear).

UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA 94720, USA

Problems in Harmonic Analysis Related to Curvature and Oscillatory Integrals

E. M. STEIN

Introduction. A series of developments in harmonic analysis in the last ten years have led us to recognize the increasing importance of certain primitive geometrical ideas such as some notions of “curvature,” and certain “metrics” related to the analysis of vector fields and nilpotent groups. Often the exploitation of these properties is intimately connected with oscillatory integrals. These developments have a close relation, both in terms of motivations and applications, with such areas of mathematics as analysis on semisimple Lie groups and symmetric spaces, several complex variables, partial differential equations, and the theory of Fourier integral operators. Because of obvious limitations of space we shall say little about these relations and applications,¹ but instead we shall concentrate on surveying the developments linking curvature, oscillatory integrals, nilpotent groups and real-variable theory.

The objects of study. We shall identify three interrelated analytical constructs, which for the purposes of this survey may be considered the central objects of study. These are: “averages of functions,” “singular integrals,” and “oscillatory integrals.” Each of these has its own long and rich history, which we will not try to give; what does interest us here are the closer connections and new points of view which have developed about them. A few introductory words about each:

Averages of functions. A basic example occurring at the earliest stages of analysis is that of the fundamental theorem of calculus. Other classical examples arise in the solution of the heat equation $\partial^2 u / \partial x^2 = \partial u / \partial t$; also the Poisson integral solving the Dirichlet problem, etc. The basic real-variable properties of such averages were subsumed in the 1930s by the Hardy-Littlewood-Wiener maximal function defined in \mathbf{R}^n by

$$(Mf)(x) = \sup_r \frac{1}{v_n r^n} \left| \int_{|y| \leq r} f(x-y) dy \right|. \quad (0.1)$$

For us the most important change in perspective that occurred is that now one realizes the increasing interest of taking averages over lower-dimensional

¹See, however, some of the literature cited below.

varieties. E.g., in (0.1) one replaces the balls of radius r by a suitable family of lower-dimensional varieties. The study of such averages is of interest not only for its own sake, but because it is relevant to other problems, such as the behavior of the classical average (0.1) as $n \rightarrow \infty$, behavior of solutions of wave equations, etc.

Singular integrals. The classical singular integral operator (arising from the Cauchy integral and from the theory of elliptic differential equations) may be written in the standard form

$$T(f)(x) = \int_{\mathbf{R}^n} f(y)K(x-y)dy \quad (0.2)$$

where the kernel K is homogeneous of degree $-n$, smooth away from the origin, and has vanishing mean-value. However the requirements of several complex variables and large classes of “subelliptic” equations have forced us to shift our interest to operators with kernels $K(x, y)$ (instead of $K(x-y)$) with the following features: (a) When the singularity of $K(x, y)$ is along the diagonal, it is no longer “isotropic,” but the behavior of K near the diagonal must be controlled by certain “distance” functions and their volume-forms; (b) One must also envisage the situation when the singularity is not just along the diagonal, but along a larger subvariety.

Oscillatory integrals. Oscillatory integrals are of many different types and not easily classifiable. Suffice it to say here that the example par excellence is the Fourier transform

$$f \rightarrow \int_{\mathbf{R}^n} e^{-2\pi i x \cdot \xi} f(x) dx = \hat{f}(\xi). \quad (0.3)$$

Other variants are obtained by replacing $f(x)dx$ by a fixed density (then the behavior of $\hat{f}(\xi)$ as $|\xi| \rightarrow \infty$ is the point of interest), or by replacing the phase $2\pi x \cdot \xi$ by a suitable generalization (then the boundedness properties of the resulting operator is what matters).

Several principles. We shall now formulate three broadly-stated (but vague) principles which may clarify the underlying approach in the development below. As with all assertions of this nature, these general observations can be taken in part as interpretation of results already obtained, and also in part as a heuristic guide to future research.

(i) For the study of operators whose kernels have singularities of nonclassical type, such related geometric notions as curvature, associated quasimetrics, and their volume-forms, play a basic role.

(ii) Properties of curvature are exploitable by the use of oscillatory integrals.

(iii) What is often crucial about the curvature in question is its nonvanishing of infinite order. The kind of phenomena that arises, then, can usually be modeled in \mathbf{R}^n in terms of polynomial behavior, or ultimately in terms of analysis on nilpotent Lie groups.

Organization of this survey. We have organized this survey in parallel with the logic of the above general principles. Thus in Part I we consider the situation in \mathbf{R}^n where the relation between curvature and oscillatory integrals is the clearest.

Next, in Part II we deal with the situation for nilpotent groups. This can be taken as a model for the “general” situation treated in Part III (which we refer to as the situation of “variable coefficients”). Some related results are briefly described in Part IV.

Part I. \mathbf{R}^n .

1. *Maximal spherical averages.* The most graphic example of the role played by curvature in real-variable theory and harmonic analysis occurs when we consider averages of functions—where these averages are not taken over solid balls as in the classical theory ((0.1) above) but, instead, where the averages are taken over the boundaries of those sets; or alternatively, one considers what happens to the standard averages in \mathbf{R}^n , but when $n \rightarrow \infty$.

For an appropriate function f given on \mathbf{R}^n , we define the mean-value of f over the sphere centered at x of radius t by

$$M_t(f)(x) = \int_{|y|=1} f(x - ty) d\sigma(y). \quad (1.1)$$

(Here $d\sigma$ denotes the normalized uniform measure on the unit sphere.)

The controlling device in the study of $\lim_{t \rightarrow 0} M_t(f)(x)$ is the maximal function

$$\mathcal{M}(f)(x) = \sup_{t>0} |M_t f(x)|, \quad (1.2)$$

which is well defined at least when f is continuous. In this context the fundamental result is

THEOREM 1. *Suppose $n \geq 2$, and $p > n/(n-1)$. Then the a priori inequality holds:*

$$\|\mathcal{M}(f)\|_{L^p} \leq A\|f\|_{L^p}. \quad (1.3)$$

(See Stein [99], Stein and Wainger [105] for $n \geq 3$, Bourgain [4] for $n = 2$.)

REMARKS. (a) One should point out first that inequalities like (1.3) can fail utterly if we replace the sphere in definition (1.1) by other surfaces which are smooth, but for which no curvature is assumed. To see this, observe first that no nontrivial result can hold for the spherical maximal function when $n = 1$ since in this case $M_t(f)(x) = \frac{1}{2}(f(x+t) + f(x-t))$, and f need merely be unbounded near one point to defeat any L^p inequality. Similar observations show that (1.3) fails for $p \leq n/(n-1)$, for the spherical maximal function, and that (1.3) may fail for any $p < \infty$ if we merely flatten the sphere near one point.

(b) The ideas in Theorem 1 have connection with several other questions in analysis which we mention briefly. First, there is a striking analogy with the behavior of the classical Radon transform defined by

$$f \rightarrow R(f)(\sigma, t) = \int_{\langle x, \sigma \rangle = t} f(x) dx,$$

with $|\sigma| = 1$, $t > 0$, and where one obtains control of $\sup_{t>0} |R(f)(\sigma, t)|$ when $p > n/(n-1)$. However, paradoxically, this analogy holds for $n \geq 3$ only; the

Radon transform result fails when $n = 2$ in view of the Besicovitch-Kakeya set. See Marstrand [59], Falconer [27], Sturtevant [108], and Oberlin and Stein [75].

Secondly, there is a close connection with convergence to initial values for solutions of the wave equation, and more generally other hyperbolic equations. For this see Stein [99], Greenleaf [37], Ruiz [86], and Sogge [91].

The proofs of results like Theorem 1 require two sets of ideas, among others: oscillatory integrals and their relation with curvature, and the use of square functions. We now turn to a brief discussion of each.

2. *Oscillatory integrals and curvature.* Since the family of mean-value operators M_t given by (1.1) are convolutions, it is natural to analyze them by the Fourier transform. Thus $(M_t(f))^\wedge(\xi) = \hat{f}(\xi)\hat{\sigma}(t\xi)$, with

$$\hat{\sigma}(\xi) = \int_{|x|=1} e^{-2\pi i x \cdot \xi} d\sigma(x).$$

It so happens that $\hat{\sigma}$ is expressible in terms of Bessel functions (i.e., $\hat{\sigma}(\xi) = c_n |\xi|^{-(n-2)/2} J_{(n-2)/2}(2\pi|\xi|)$), from which the crucial fact of the decay of $\hat{\sigma}$ at infinity follows easily:

$$|\hat{\sigma}(\xi)| \leq A |\xi|^{-(n-1)/2}. \quad (2.1)$$

(See, e.g., Stein and Weiss [106, Chapter 4].)

It is the exploitation of this decay that makes Theorem 1 possible. To make clear the importance of curvature, let us consider the analogous question where the sphere in \mathbf{R}^n is replaced by (say) a smooth hypersurface S , and $d\sigma$ is replaced by $d\mu = \psi d\sigma$, with ψ a C_0^∞ cut-off function, and $d\sigma$ the induced Lebesgue measure on S . Define $\hat{\mu}(\xi)$ by

$$\hat{\mu}(\xi) = \int_S e^{-2\pi i x \cdot \xi} d\mu. \quad (2.2)$$

Then it is known (facts which go back to Van der Corput when $n = 2$, Hlawka [47] when $n \geq 3$; see also Herz [46], Hörmander [51]), that

$$\hat{\mu}(\xi) = O(|\xi|^{-(n-1)/2}), \quad \text{as } |\xi| \rightarrow \infty,$$

essentially if and only if S has nonvanishing Gauss curvature at each point of the support of ψ .

For future reference we now state two further results relating (some kind) of nonvanishing curvature and decay of the Fourier transform. For the first result, S^k will denote a smooth k -dimensional manifold embedded in \mathbf{R}^n , $d\mu = \psi d\sigma$, with $\psi \in C_0^\infty$, and $d\sigma$ the induced measure on S^k . We shall say that S^k is of *finite type* at a point $x_0 \in S^k$, if S^k has only a finite order of contact with any affine hyperplane through x_0 . Then one can observe the following (see Stein [103])

LEMMA 1. *With S^k and $d\mu$ as above, there exists an $\varepsilon > 0$, $\varepsilon = \varepsilon(S^k)$, so that $\hat{\mu}(\xi) = O(|\xi|^{-\varepsilon})$, as $|\xi| \rightarrow \infty$.*

This result holds in particular when S^k is real-analytic and does not lie in any affine hyperplane, as was previously shown by Björck [2].

Suppose, however, that we wish to reinstate the full decay ($O(|\xi|^{-(n-1)/2})$), at least for hypersurfaces in \mathbf{R}^n . Then this is possible if we insert a mitigating factor involving the Gaussian curvature. For our second result, S denotes an arbitrary smooth hypersurface in \mathbf{R}^n , $K(x)$ its Gaussian curvature at $x \in S$, and $d\mu = \psi d\sigma$ as before.

LEMMA 2. *For N sufficiently large,*

$$\int_S e^{-2\pi i x \cdot \xi} |K(x)|^N d\sigma(x) = O(|\xi|^{-(n-1)/2}). \quad (2.3)$$

In fact (2.3) holds when $N \geq 2n - 2$.

This lemma is in Sogge and Stein [92]. Another type of estimate for (2.2) which gives $O(|\xi|^{-(n-1)/2})$, as $|\xi| \rightarrow \infty$, in “most” directions can be found in Randol [80] and Svenson [112]. See also the estimate in §16, which involves certain other geometric notions of interest. In connection with Lemma 2 we state the following

PROBLEM. (a) What is the least N for which (2.3) holds?

(b) More interestingly, what is the “largest” function (in place of $|K(x)|^N$) for which the integral (2.3) is $O(|\xi|^{-(n-1)/2})$?

3. *Square functions.* The decay properties of the Fourier transform of a surface-carried measure can be exploited by the use of certain square functions. The simplest version of these which arises when considering spherical means is the one given by

$$S(f)(x) = \left(\int_0^\infty \left| \frac{dM_t(f)(x)}{dt} \right|^2 t dt \right)^{1/2}. \quad (3.1)$$

By Plancherel’s formula, and a decay property like (2.1), it can be shown that $\|S(f)\|_{L^2} \leq c\|f\|_{L^2}$, when $n \geq 4$. Then since

$$t^n M_t = \int_0^t \frac{d(s^n M_s)}{ds} ds = n \int_0^t M_s s^{n-1} ds + \int_0^t \frac{dM_s}{ds} s^n ds$$

it follows easily that $M(f) \leq c\{S(f) + M(f)\}$, which shows how the case $p = 2$, $n \geq 4$ of Theorem 1 can be proved. More refined versions of this argument are needed to deal with the situation $p = 2$, $n \geq 3$. An example of what can be achieved by such refinements is as follows. Suppose $m(\xi)$ is a C^1 function on \mathbf{R}^n which satisfies, for some $\varepsilon > 0$,

$$|m(\xi)| + |\nabla m(\xi)| \leq c(1 + |\xi|)^{-1/2-\varepsilon}. \quad (3.2)$$

Define $M_t f$ by $M_t(f)^\sim(\xi) = m(t\xi)\hat{f}(\xi)$.

LEMMA. $\|\sup_{t>0} |M_t(f)|\|_{L^2} \leq A\|f\|_{L^2}$.

This version is in Sogge and Stein [92]. Other versions are in Bourgain [3], Carbery [10], and Rubio de Francia [85].

The result above is sharp in the sense that it fails when $\varepsilon = 0$.

The L^p theory $p \neq 2$ is actually implicit in this approach to the L^2 theory via square functions. In fact an examination of the above shows that there is some slack in the argument, and a “larger” or rougher version of the operator M is still bounded on L^2 . Using the L^2 credit just established we can combine it with the fact that a “better” version of M can actually be treated (for all L^p) by standard means. Note, however, that the case $n = 2$ is more intricate here. As Bourgain shows, the standard L^p estimates have to be replaced by some tricky geometric arguments.

Thus the general scheme for proving maximal L^p inequalities (to which we will return again) can be summarized: first use square functions (for L^2); then interpolate (for L^p). The paradigm for this argument was actually developed in the more abstract context of a general maximal theorem for symmetric diffusion one-parameter semi-groups. (See Stein [95], [96].) The application of square functions in one form or another is crucial for several other results we will discuss. See also §9 below.

4. *Further results for surface averages.* Having outlined the ideas in the proof of Theorem 1, we come now to some variants and consequences. The first deals with a generalization of M_t where the sphere is replaced by a more general smooth hypersurface S . If we define $d\mu(x) = \psi(x) d\sigma$ as in (2.2) we can write

$$M_t(f)(x) = \int_S f(x - ty) d\mu(y), \quad t > 0,$$

and $M(f) = \sup_{t>0} |M_t(f)|$.

THEOREM 2. *Suppose that the Gauss curvature of S does not vanish of infinite order at any point of S . Then there exists a $p_0 = p_0(S)$, $p_0 < \infty$, so that*

$$\|M(f)\|_{L^p} \leq A_p \|f\|_{L^p} \quad \text{whenever } p_0 < p \leq \infty. \quad (4.1)$$

Observe that the assumptions are verified if S is a compact real-analytic hypersurface. The proof of this theorem follows from the decay estimate (2.3) and an interpolation argument with L^∞ (where the nonvanishing of infinite order is exploited). Note that allowing the curvature of S to vanish of infinite order at one point may invalidate (4.1) for all $p < \infty$. For details of the proof of Theorem 2, see Sogge and Stein [92], for $n \geq 3$, and Bourgain [5] for $n = 2$. Earlier relevant ideas are in Greenleaf [37], and Cowling and Mauceri [21]; see also Cowling and Mauceri [22].

In this connection it is natural to raise the following question:

PROBLEM. (a) It may be conjectured that the maximal inequality (4.1) holds for a nontrivial range of exponents if and only if at each point, S makes a contact of finite order with its tangent hyperplane.

(b) Related to this is the problem of finding the smallest value of $p_0 = p_0(S)$ for which (4.1) holds.

5. *Results for \mathbf{R}^n , $n \rightarrow \infty$.* The results and techniques described above in connection with the nonorthodox maximal functions apply also to the usual

maximal functions and other standard operators in harmonic analysis in \mathbf{R}^n , in studying the question of what happens to the bounds of the operators when we let $n \rightarrow \infty$. We begin with the usual (centered) maximal function defined by

$$M^{(n)}(f)(x) = \sup_{r>0} \frac{1}{v_n r^n} \left| \int_{|y| \leq r} f(x-y) dy \right| \quad (5.1)$$

with v_n the volume of the unit ball.

THEOREM 3. (a) $\|M^{(n)}(f)\|_p \leq A_p \|f\|_p$, $1 < p \leq \infty$, where the constant A_p may be taken to be independent of n .

(b) For the weak-type inequality we can assert that

$$m\{x | M^{(n)}f(x) > \alpha\} \leq c_n/\alpha \|f\|_{L^1}, \quad \text{with } c_n = O(n).$$

The idea of the proof of (a) is that we can transfer the spherical maximal inequality (1.3), for a given p and a fixed dimension, to higher dimensions, without increasing the bound. (This transference does not work for the standard maximal function!) Part (b) requires a separate argument which utilizes the general maximal ergodic theorem. Further details are in Stein and Stromberg [104]; see also Stein [100–102].

Another result of this type deals with the basic singular integrals, the Riesz transforms, R_j , $j = 1, \dots, n$, defined by $R_j(f)^\wedge(\xi) = (i\xi_j/|\xi|)\hat{f}(\xi)$. In analogy with the above one can deal with the problem for the bounds for $|R(f)| = (\sum_{j=1}^n |R_j(f)|^2)^{1/2}$ in terms of appropriate square functions—in this case, the n -dimensional extensions of the g -functions of Littlewood and Paley.

The functions in question are

$$g_1(f)(x) = \left(\int_0^\infty \left| \frac{\partial u}{\partial t}(x, t) \right|^2 t dt \right)^{1/2}, \quad g(f)(x) = \left(\int_0^\infty |\nabla u(x, t)|^2 t dt \right)^{1/2}$$

with $u(x, t)$ the Poisson integral of f .

PROPOSITION. When $1 < p \leq \infty$, the L^p norms of f , $R(f)$, $g(f)$, and $g_1(f)$ are all equivalent, with bounds that can be taken independent of n .

The equivalence of norms (with bounds independent of the dimension) of f , $g(f)$, and $g_1(f)$ are implicit in the treatment given for these square functions in Stein [96] and [97]; see also P. A. Meyer [62]. The inequalities involving the Riesz transforms follow because of the pointwise estimate

$$g_1(Rf)(x) \leq g(f)(x). \quad (5.3)$$

(A different approach to the bounds of $R(f)$, which exploits the “method of rotations” has been developed by Duoandikoetxea and Rubio de Francia [25].)

After this digression we return to Theorem 3 and ask whether centered balls can be replaced by dilates of a fixed convex symmetric body B . That is, for B fixed we consider M_B defined by

$$(M_B f)(x) = \sup_{r>0} \frac{1}{m(rB)} \left| \int_{rB} f(x-y) dy \right|$$

where $rB = \{rx, x \in B\}$, and $m(rB)$ is its volume. We raise the question of whether or not we can make L^p estimates for M_B which are independent of B and the dimension n .

THEOREM 4. (a) $\|M_B(f)\|_p \leq A_p \|f\|_p$, if $\frac{3}{2} < p \leq \infty$, with A_p independent of B and n .

(b) $m\{x | M_B(f)(x) > \alpha\} \leq \bar{c}_n / \alpha \|f\|_1$, with $\bar{c}_n = O(n \log n)$, and \bar{c}_n otherwise independent of B .

The case $p \geq 2$ in (a) above was proved by Bourgain [3] and was based on his key lemma (below). Using this lemma both Bourgain [6] and Carbery [10] independently extended the result to $p > \frac{3}{2}$. (The role of the special exponent $\frac{3}{2}$ will be clarified momentarily.) Part (b) of the theorem, which is proved by very different methods, is in Stein and Stromberg [104].

The main idea of the proof of (a) is the following remarkable universal decay estimate for the Fourier transform of the characteristic function χ_B , of a convex symmetric body B in \mathbf{R}^n of unit volume. Let $\hat{\chi}_B(\xi) = \int_B e^{-2\pi i x \cdot \xi} dx$.

LEMMA. *There exists an $L \in \text{GL}(n, \mathbf{R})$, $L = L(B)$, so that if $m(\xi) = \hat{\chi}_B(L(\xi))$, then*

$$\left. \begin{aligned} |m(\xi)| &\leq c/|\xi|, \\ |m(\xi) - 1| &\leq c|\xi|, \\ |\langle \xi, \nabla m(\xi) \rangle| &\leq c, \end{aligned} \right\} \quad (5.4)$$

where c is a universal constant.

Once the lemma is proved, the theorem for $p \geq 2$ follows easily by a variant of the lemma in §3. For the value of p less than 2 we notice that the estimates (5.4) are the same as one can make for the Fourier transform of the surface-carried measure on the unit sphere in \mathbf{R}^3 ! Since in this case (Theorem 1) the result fails for $p \leq \frac{3}{2}$, we cannot expect to succeed here beyond $\frac{3}{2}$ with the estimate (5.4) given by the lemma. However, in the other direction, this observation shows that there is room to spare in the L^2 estimates, and one can hope, as in the case of the spherical maximal function, to achieve $p > \frac{3}{2}$. The execution of this requires a bootstrap argument involving interpolation, which has some similarity to the proof for the "lacunary maximal functions" (in §14) and the "flat case" for curves (in §15). At this stage it may be worthwhile to record the following problem.

PROBLEM. (a) Do $M^{(n)}$ (and for that matter M_B) have weak-type bounds independent of n ?

(b) Do the Riesz transforms have weak-type bounds independent of n ?

(c) Can part (a) of Theorem 4 be extended to the range $1 < p \leq \frac{3}{2}$?

6. *Where the subvariety is fixed.* In §§1 and 4 we dealt with averages over dilates of a fixed surface, that is, when the averages are taken over sets of the form $\{x + \varepsilon S\}_{\varepsilon > 0}$; we saw how curvature properties played a crucial role and were exploitable by oscillatory integrals and square functions. Here we shall deal with the case when the objects of interest—maximal functions, singular integrals,

etc.—are taken with respect to a fixed submanifold. There is some analogy to matters described above, but also important differences, as we shall see. The theory for the fixed submanifold began with the case when the submanifold was a curve. It flourished with the early work of Nagel, Riviere, and Wainger [66, and 67]; it was initially developed under the twin incentives of coming to grips with the “method of rotations” for nonisotropic singular integrals and better understanding the boundary behavior of Poisson integral on symmetric spaces. (See the account in Stein and Wainger [105]; some earlier motivation is in Stein [98].)

We shall state our results in two forms. First in their “global versions,” and then in their “local” versions. The global theorems are, implicitly, consequences of the local versions; their real role is to serve as models for the general theorems. We have formulated the global theorems separately to emphasize a principle formulated earlier, namely, that the properties related to curvature not vanishing of infinite order should be modeled by polynomial behavior.

Fix a polynomial mapping $P: \mathbf{R}^k \rightarrow \mathbf{R}^n$, and let K be a standard Calderón-Zygmund kernel in \mathbf{R}^k , i.e., K is homogeneous of degree $-k$, smooth away from the origin, and has vanishing mean-value. (Here the image of P in \mathbf{R}^n plays the role of our submanifold.) We then define the singular integral T acting on appropriate functions on \mathbf{R}^n by

$$T(f)(x) = \text{p. v.} \int_{\mathbf{R}^k} f(x - P(u))K(u) du. \quad (6.1)$$

Similarly we define the maximal function M by

$$M(f)(x) = \sup_{0 < r < \infty} r^{-k} \left| \int_{|u| < r} f(x - P(u)) du \right|. \quad (6.2)$$

In order to formulate the local versions we replace the polynomial mapping P by a smooth embedding ρ of the unit ball B_1 in \mathbf{R}^k , into \mathbf{R}^n ; i.e., $\mathbf{R}^k \supset B_1: \rho \rightarrow \mathbf{R}^n$. We also suppose that $\rho(0) = 0$. The key assumption to make is that $\rho(B_1)$ is of finite type (in the sense of §2) at the origin.

We then define, in analogy with the above,

$$T_1(f)(x) = \text{p. v.} \int_{|u| \leq 1} f(x - \rho(u))K(u) du, \quad (6.1')$$

$$M_1(f)(x) = \sup_{0 < r \leq 1} r^{-k} \left| \int_{|u| \leq r} f(x - \rho(u)) du \right|. \quad (6.2')$$

THEOREM 5. (a) T is bounded on L^p , $1 < p < \infty$; the bounds for the operator do not depend on the coefficients of the polynomial P , but only on the total degree of P .

(b) T_1 is bounded on L^p , $1 < p < \infty$, if we assume ρ is of finite type at the origin.

There are similar results for M and M_1 in the range $1 < p \leq \infty$.

Extensions and variants of this result in the context of nilpotent Lie groups are in §8 below.

REMARKS. Some degree of curvature (in the guise of a “finite type” condition or an attenuated convexity) seems necessary for the conclusions of part (b) of Theorem 5. In fact, there is an example of a smooth curve in the plane—which is infinitely flat at the origin—for which the inequalities fail for all $p < \infty$. However, in marked distinction with the situation in §§1 and 4, there are (besides the straight-line) curves which are infinitely flat at the origin for which there are positive results for all p . Further details are in §15 below.

We sketch briefly some of the ideas of the proof of Theorem 5. The fact that the local maximal operator M_1 is bounded on L^2 can be seen by using a square function argument, not unlike that in §3. To define the requisite square function we proceed as follows. Let $\rho(u) \sim \sum a_\alpha u^\alpha$ be the Taylor development of ρ at the origin; here a_α are vectors in \mathbf{R}^k . Using these vectors we can find disjoint (possibly empty) subspaces V_1, V_2, \dots of \mathbf{R}^k so that $V_1 + V_2 + \dots + V_m$ is spanned by $\{a_\alpha\}_{|\alpha| \leq m}$. Since ρ is of finite type at the origin, $\sum_{j=1}^N V_j = \mathbf{R}^k$ for some N . Next define dilations δ_t by $\delta_t(v) = t^j v$, $v \in V_j$, $t > 0$. Then $\delta_t^{-1} \rho(tu) \rightarrow \rho_0(u)$ as $t \rightarrow 0$, where ρ_0 is a polynomial, which again is of finite type at the origin. Next choose $\phi \in C_0^\infty(\mathbf{R}^k)$, $\psi \in C_0^\infty(\mathbf{R}^n)$, $\phi \geq 0$, $\psi \geq 0$, with $\int_{\mathbf{R}^k} \phi du = \int_{\mathbf{R}^n} \psi dy = 1$. Set

$$M^j(f) = \int_{\mathbf{R}^k} f(x - \rho(2^{-j}u))\phi(u) du, \quad N^j(f) = \int_{\mathbf{R}^n} f(x - \delta_{2^{-j}}(y))\psi(y) dy.$$

The square function in question is then given by

$$S(f)(x) = \left(\sum_{j \geq 0} |M^j(f)(x) - N^j(f)(x)|^2 \right)^{1/2}.$$

The fact that $\|S(f)\|_{L^2} \leq A\|f\|_{L^2}$ follows easily by appealing to (a slight variant of) the decay estimate given by Lemma 1 in §2. The L^p theory for M_1 can then be carried out in the same spirit as the case of the curve: What is important here is that there is an “ ε slack” in Lemma 1 which can be exploited.

It should be pointed out that the L^2 result for the operator T in (6.1) is equivalent, via Plancherel’s theorem, to the following estimate for oscillatory integrals: let P_0 denote an arbitrary real-valued polynomial on \mathbf{R}^k ; then

$$\left| \text{p.v.} \int_{\mathbf{R}^k} e^{iP_0(u)} K(u) du \right| \leq A \quad (6.3)$$

where the constant may be taken to depend only on the degree of P_0 , and not otherwise on the coefficients. For this, see Stein [103].

Part II. Nilpotent Lie groups.

7. *Nilpotent Lie groups: background.* Let us briefly describe some of the standard aspects of analysis on nilpotent groups that are relevant here.²

Let N be a simply-connected nilpotent Lie group, which we will always identify with its Lie algebra \mathfrak{n} via the exponential map. We assume a fixed direct sum

²In addition to the papers cited below the reader may consult the survey of Goodman [36].

decomposition of $\mathfrak{n} = \sum_{j=1}^m \mathfrak{n}_j$, and a sequence of strictly positive numbers a_1, \dots, a_m , with which we define dilations $\{\delta_t\}$, $t > 0$, on \mathfrak{n} by the formula $\delta_t(v) = t^{a_j}v$, if $v \in \mathfrak{n}_j$. In general such dilations are not automorphisms of the algebra \mathfrak{n} ; however, when they are we will refer to them as *automorphic dilations*. The group N together with these dilations will be called a homogeneous group. An example is the Heisenberg group H^1 , identified with $\mathbf{R}^3 = \{(x, y, z)\}$; its multiplication law is

$$(x, y, z) \cdot (x', y', z') = (x + x', y + y', z + z' + x'y - xy'). \quad (7.1)$$

We mention here two different types of dilations on H^1 :

$$\begin{aligned} (a) \quad & \delta_t: (x, y, z) \rightarrow (tx, ty, t^2z), \\ (b) \quad & \delta_t: (x, y, z) \rightarrow (tx, ty, tz). \end{aligned} \quad (7.2)$$

The first dilations are automorphic; they are intimately connected with the geometry of H^1 induced by its realization as the boundary of a domain in \mathbf{C}^2 . The second dilations (which are not automorphisms) may be said to be the vestiges remaining from elliptic theory, insofar as it applies in view of the $\bar{\partial}$ -Neumann problem for this domain.

We begin by concentrating on homogeneous nilpotent groups. The more general situation will be taken up in §8. Incidentally, the groups that arise in the examples cited after (7.3) are all homogeneous.

The structure of a homogeneous group naturally leads to some basic geometric constructs: a (left-invariant) quasidistance $\rho(x, y)$ which is homogeneous under dilations, and defined by $\rho(x, y) = |y^{-1} \cdot x|$, where $|\cdot|$ is a suitable homogeneous norm function on N ; also the volume function $V(t)$, giving the volume of the ball of radius t (with distance measured by ρ). Because of homogeneity, $V(t) = ct^Q$, $t > 0$, where Q is the homogeneous dimension of N .

We can then define the “standard singular integrals” on N in close analogy with the classical case of \mathbf{R}^n (as described in (0.2)). Thus we consider convolution operators $T: f \rightarrow f * K$, where the principal-value kernel K satisfies

$$\begin{aligned} (a) \quad & K \text{ is smooth away from the origin,} \\ (b) \quad & K(\delta_t x) = t^{-Q} K(x), \quad t > 0, \\ (c) \quad & K \text{ has vanishing mean-value.} \end{aligned} \quad (7.3)$$

Examples where such operators occur are: (1) Intertwining operators (see Knapp and Stein [54]); (2) The Cauchy-Szegő operator for the generalized upper half space (see Koranyi and Vagi [55]); (3) Operators arising for the critical estimates for \square_b , the Kohn-Laplacian (see Folland and Stein [34]); and, more generally, in the analysis of the Hörmander [48] operators, as studied in Rothschild and Stein [84].

However it is not our intention to dwell here on these “standard” operators, which are by now well understood. Rather we turn to operators on nilpotent groups whose kernels are carried on lower-dimensional submanifolds in analogy with results in Part I. We are in fact forced to consider these generalizations

when we make a detailed analysis of the $\bar{\partial}$ -Neumann. The reason for this is that the explicit formulae that arise (see Henkin [45], Lieb [57], and also the survey of Krantz [56] for the operators solving $\bar{\partial}$; for the $\bar{\partial}$ -Neumann problem see Phong [76], Harvey and Polking [43], Lieb and Range [58], Stanton [94], and the survey of Beals, C. Fefferman, and Grossman [1]) lead to kernels which are products of the two kinds of homogeneities (7.2). How do we come to grips with these two conflicting homogeneities? It turns out that the kernel

$$\tilde{K}(x, y, z) = L(x, y)\Delta(z) \quad (7.4)$$

on H^1 , where $L(x, y)$ is a standard Calderón-Zygmund kernel on \mathbf{R}^2 , and $\Delta(z)$ is the Dirac delta-function, is simultaneously homogeneous with respect to both types of dilations, each of the critical degree! So it is not surprising that the analysis of the operator \tilde{T} , given by $\tilde{T}f = f * \tilde{K}$, should be important for the $\bar{\partial}$ -Neumann problem.

The L^2 theory of \tilde{T} is (because of the particular form (7.4) of \tilde{K} and the multiplication law (7.1)) most easily treated by taking the Fourier transform in the z -variable. This reduces the L^2 theory of \tilde{T} to that of certain oscillatory operators which, put in somewhat more general form, are as follows: Fix m , let $B(\cdot, \cdot)$ be a real-valued bilinear form on $\mathbf{R}^m \times \mathbf{R}^m$, K a standard Calderón-Zygmund kernel on \mathbf{R}^m , and consider the operator T given by

$$(Tf)(x) = \text{p. v.} \int_{\mathbf{R}^m} e^{iB(x, y)} K(x - y) f(y) dy. \quad (7.5)$$

For the Heisenberg group \mathbf{H}^n we take $m = 2n$, $B(x, y) = \lambda \langle x, y \rangle$, where $\langle \cdot, \cdot \rangle$ is the symplectic form on $\mathbf{R}^{2n} \times \mathbf{R}^{2n}$ which defines the multiplication laws on \mathbf{H}^n , with $-\infty < \lambda < \infty$. In this case the operators (7.5) reduce to “twisted convolution” operators on the Heisenberg group—operators which have an interest in their own right (see, e.g., Segal [89], A. Grossman, Lempert, and Stein [40], Howe [52], Mauceri, Piccardello, and Ricci [60]).

To summarize this background section: In analogy with Part I we have isolated certain convolution operators carried on lower-dimensional submanifolds and the oscillatory integrals they lead to; the curvature condition in this context is subsumed in the fact that $\langle \cdot, \cdot \rangle$ is nondegenerate. A more geometric description of this curvature is as follows: for each point $P = (x, y, z)$ in H^1 we associate the two-dimensional (affine) plane M_P , where M_P is the left group translate of M_0 by P , with $M_0 = \{(x, y, 0)\}$. The value of $\tilde{T}(f)$ at P is given by an integration over M_P ; the curvature here is the “rotation” of the plane M_P as P varies. The theory of the operators \tilde{T} and (7.5) can be found in Geller and Stein [35] and Phong and Stein [77–79]. Further related results are in §§8 and 11 below.

8. *Some general results for nilpotent Lie groups.* We now come to a series of results which will extend and unify much of the theory in \mathbf{R}^n , presented in §6, with the ideas for nilpotent Lie groups discussed in §7.

Here N will be an arbitrary simply-connected nilpotent group, and $\{\delta_t\}$, $t > 0$, a fixed family of dilations, not necessarily automorphic. We fix a real-analytic submanifold V of $N/\{0\}$ which, for simplicity, we assume connected, and also

that V is homogeneous in the sense that $\delta_t(V) = V$, all $t > 0$. We shall assume we are given a distribution K on N which is invariant under dilations in the sense that $\langle K, f(x) \rangle = \langle K, f(\delta_t(x)) \rangle$, all $t > 0$, and which agrees, away from the origin, with a measure on V with smooth density (with respect to induced Lebesgue measure $d\sigma$ on V), i.e., of the form $\psi(x) d\sigma(x)$, with ψ smooth and having compact support when restricted to $\{|x| = 1\} \cap V$. We define T by

$$Tf = f * K. \quad (8.1)$$

In the same setting we can consider an analogous maximal function. Let $d\mu = \chi_\Gamma d\sigma$, where Γ is the characteristic function of a compact subset Γ of V . We define $d\mu_t$ by $\langle d\mu_t, f(x) \rangle = \langle d\mu, f(\delta_t x) \rangle$, $t > 0$. Set

$$(Mf)(x) = \sup_{t>0} |(f * d\mu_t)(x)|. \quad (8.2)$$

THEOREM 6. *Both T and M , defined by (8.1) and (8.2), are bounded on $L^p(N)$ to itself, $1 < p$.*

A special case of this result deserves to be formulated separately. Suppose K is a principal-value kernel, supported on all of N , which satisfies (7.3). In this case we set $Tf = f * K$, and similarly if Γ is an open subset of N of compact closure we write

$$(Mf)(x) = \sup_{t>0} \left| \int_{\Gamma} f(x \cdot (\delta_t(y))^{-1}) dy \right|. \quad (8.3)$$

COROLLARY 1. *T and M so defined are both bounded on $L^p(N)$, $1 < p$.*

This shows that the assumption that the group be equipped with automorphic dilations, previously considered essential in proving such results, is in reality unnecessary. Notice however that the balls which are implicit in the definition of (8.3) do not satisfy the Vitali covering property (i.e., if N is not a homogeneous group, then $|y^{-1} \cdot x| = \rho(x, y)$ does not satisfy the quasitriangle inequality). Thus the standard methods do not apply here.

Another corollary of the proof is a generalization of Theorem 5, part (a). For this purpose let P be a polynomial mapping from \mathbf{R}^k to N , and K a standard Calderón-Zygmund kernel on \mathbf{R}^k .

COROLLARY 2. *The operators*

$$f \rightarrow \text{p. v.} \int_{\mathbf{R}^k} f(x \cdot (P(u))^{-1}) K(u) du$$

and

$$f \rightarrow \sup_{0 < r} r^{-k} \left| \int_{|u| \leq r} f(x \cdot (P(u))^{-1}) du \right|$$

are bounded on $L^p(N)$, $1 < p < \infty$. Their bounds may be taken not to depend on the coefficients of P , but only on its degree.

A final corollary provides a unification with the oscillatory integrals in (6.3) and (7.5) and Corollary 1.

COROLLARY 3. *Let $P(x, y)$ be a fixed, real polynomial on $N \times N$. Then the operator*

$$f \rightarrow \text{p. v.} \int_N e^{iP(x, y)} K(y^{-1} \cdot x) f(y) dy \quad (8.4)$$

is bounded on $L^p(N)$ to itself, $1 < p < \infty$; again the bounds depend (insofar as P is concerned) only on its total degree.

Besides the literature already cited, treatment of some cases of operators (8.1), (8.2), and (8.4) may be found in Strichartz [109], Müller [63, 64], Christ [14], and Greenleaf [39]. The general theorem as stated, and its corollaries, are contained in the recent work of Ricci and the author [81–83].

We shall now briefly describe the three main ideas involved in the proof of the theorem. First, one passes from the nilpotent group N with its nonautomorphic dilations δ_t , to a larger nilpotent group G which has corresponding automorphic dilations, so that N is a quotient of G . In fact we take G to be the “free” nilpotent group of sufficiently high step whose Lie algebra \mathfrak{g} has as its generators a basis of \mathfrak{n} . The boundedness properties of the operators T and M may then be reduced to the boundedness properties of (approximately constructed) corresponding operators T' and M' on G . This procedure is in reality a streamlined form of the “lifting” method which will be discussed again in §10.

Secondly, the “curvature” properties of our manifold V are expressible by the fact that it is real-analytic and may be assumed to generate the group N . The following lemma takes advantage of this situation.

LEMMA. *Suppose V is a k -dimensional real-analytic submanifold in a connected Lie group G ; suppose that V generates G . Let $d\mu$ be a measure supported in V with a smooth, compactly supported density. Then $d\nu = d\mu * d\mu * \cdots * d\mu$ ($2^{\dim(G)-k}$ factors), as a measure on G , is absolutely continuous with respect to Haar measure, $d\nu = h dx$, where h satisfies an L^1 modulus of continuity with ε -Hölder exponent, $\varepsilon > 0$.*

The third idea is to use appropriate orthogonality arguments. We now turn to a further discussion of this.

9. *Square functions and orthogonality.* Three variations on the theme of orthogonality:

(i) *The method of TT^* .* This phrase is meant to convey the truism that to prove the boundedness of T on L^2 it suffices to do the same for TT^* (or T^*T). The reason this observation is useful in applications is that if T is represented by the kernel K , $(Tf)(x) = \int K(x, y)f(y) dy$, then TT^* is represented by the kernel

$$\int K(x, z)\overline{K}(y, z) dz, \quad (9.1)$$

and this kernel is often better than K because the integration in (9.1) can have both a smoothing effect and can take into account cancellation properties of K . Classical examples arise when dealing with unitary operators such as the Fourier transform, the Hilbert transform, etc. A trickier example is the theorem

of Kolmogorov-Seliverstov-Plessner (see Zygmund [115]) which shows, incidentally, that the computations which prove the boundedness of TT^* may be quite different from those that show it for T^*T . Other examples arise in the theory of semigroups (see Stein [96]) and for Fourier integral operators (see Hörmander [49]); closer at hand are the oscillatory integrals (7.5), (8.4), and (11.4) and (12.3) below, and the treatment of Hilbert transforms and maximal functions for “variable curves” (discussed further in §11), as carried out in Nagel, Stein, and Wainger [69].

(ii) *Almost orthogonality.* The idea of TT^* does not always suffice to do the job. It can be refined as follows: One has boundedness of T in L^2 if we can find a decomposition of T as $\sum T_j$, with the norms of the T_j uniformly bounded, and an approximate mutual orthogonality of the different pieces T_j ; the latter is expressed by the requirement that the norms of both $T_i T_j^*$ and $T_i^* T_j$ tend to zero suitably when $|i - j| \rightarrow \infty$. This is the lemma of almost orthogonality of Cotlar and the author. It was applied to prove the L^2 boundedness of the standard singular integrals on nilpotent groups discussed above (see Knapp and Stein [54]), and has since found many applications and modifications. One such variant of relevance here (due to Christ [14], and in another form to Greenleaf [38]) is the observation that since the norms of the T_j are bounded, the control of $T_i T_j^*$ can be reduced (by applying the principle of TT^* repeatedly) to the control of $(T_i^* T_i)^k T_j^*$, for any fixed (large) integer k . This is of particular use in conjunction with arguments such as the lemma in §8 above.

(iii) *More square functions.* We have already seen several variants of square functions (in §§3 and 4) and discussed their usefulness in proving L^p inequalities. Here we shall describe a version closely related to the almost-orthogonal decompositions above. Suppose we can write $T = \sum T_j$, with T_j factorable as $T_j = B_j^* A_j$. We then define two square functions S_1 and S_2 : $S_1(f) = (\sum_j |A_j f|^2)^{1/2}$, $S_2(f) = (\sum_j |B_j f|^2)^{1/2}$. It then follows that the boundedness of T on L^p is a direct consequence of the boundedness of $f \rightarrow S_1(f)$ and $g \rightarrow S_2(g)$, on L^p and $L^{p'}$ respectively. This type of decomposition and use of square functions plays an important role in the theory of singular integrals of Coifman, MacIntosh, and Meyer [16], and David and Journé [24]; the reader should also consult David [23].

Part III. Variable coefficients.

10. *Geometry and analysis of vector fields.* The passage to variable coefficients via nilpotent groups is nowhere better illustrated than in the study of analysis and geometry of vector fields. The idea of using nilpotent groups was initiated in the study of $\bar{\partial}_b$ and the Kohn Laplacian, and then extended to deal with the operators analyzed by Hörmander [48]. Let us describe it briefly.

Let X_1, \dots, X_m be m smooth real vector fields in (a neighborhood of the origin of) \mathbf{R}^n . We assume that these vector fields and their commutators of order not greater than k span near the origin. It is then possible to “lift” these vector fields by placing them in a larger space, $\mathbf{R}^n \times \mathbf{R}^{n'}$ (with additional variables

$t \in \mathbf{R}^{n'}$) so that the lifted vector fields are of the form

$$\tilde{X}_i = X_i + \sum_{j=1}^{n'} a_{ij}(x, t) \frac{\partial}{\partial t_j}, \quad i = 1, \dots, m, \quad (10.1)$$

and so that the \tilde{X}_i have the twin properties: (i) The \tilde{X}_i and their commutators of length not greater than k span near the origin in $\mathbf{R}^n \times \mathbf{R}^{n'}$; (ii) The \tilde{X}_i are “free,” that is, these vector fields and their commutators satisfy the same relations that the elements of the free nilpotent Lie algebra $\mathcal{F}_{m,k}$ satisfy. (The Lie algebra is obtained as the quotient of the free Lie algebra of m generators Y_1, Y_2, \dots, Y_m , divided by the ideal generated by all commutators of length $> k$.)

Besides this lifting there is a basic approximation property. First write the vector fields Y_1, \dots, Y_m in canonical coordinates $(y_1, \dots, y_{n+n'})$

$$Y_i = \sum_j b_{ij}(y) \partial / \partial y_j.$$

Then for each fixed $(x_0, t_0) \in \mathbf{R}^n \times \mathbf{R}^{n'}$ near the origin we can find a coordinate system $(y_1, \dots, y_{n+n'})$ centered at (x_0, t_0) so that

$$\tilde{X}_j = Y_j + R_j, \quad j = 1, \dots, m, \quad (10.2)$$

with R_j vanishing of appropriately high order as $(x, t) \rightarrow (x_0, t_0)$ (i.e., as $(y_1, \dots, y_{n+n'}) \rightarrow 0$).

The above two-fold procedure allows one to reduce the study of, e.g., the operator $\sum X_j^2$ to that of $\sum \tilde{X}_j^2$, which in turn is modeled on the left-invariant (and homogeneous) operator $\sum Y_j^2$ on the nilpotent group G whose Lie algebra is $\mathcal{F}_{m,k}$. This procedure of appealing to analysis on nilpotent Lie groups goes back to Folland and Stein [34] and Rothschild and Stein [84]; it has since undergone substantial development. (In this connection see, e.g., the survey of Helffer and Nourigatt [44].)

In addition one may seek a better understanding of the geometry of the vector fields, also, for example, the size of the fundamental solution of the operator $\sum X_j^2$. To do this two basic notions are needed: an appropriate “metric” (or quasidistance) and the volume formula for the resulting balls.

Given the vector fields X_1, \dots, X_m , we can then define a naturally induced metric: the distance $\rho(x, y)$ is the least time it takes to travel from x to y along a curve pointing in the directions of X_1, \dots, X_m , always moving at “unit speed.” More precisely, $\rho(x, y) = \infimum$ of T so that $\gamma(0) = x$, $\gamma(T) = y$, where $\gamma(t)$, $0 \leq t \leq T$, is a rectifiable curve so that $\dot{\gamma}(t) = \sum_{j=1}^m a_j(t) X_j(\gamma(t))$, with $|a_j(t)| \leq 1$.

In order to work with this distance it is quite useful to describe it in several equivalent forms. We shall limit ourselves here to one such form, linking it with the exponential mapping, and we shall also describe a formula for $V_x(\delta) = \text{volume of the ball } \{y | \rho(x, y) < \delta\}$.

For each ordered l -tuple of integers $I = (i_1, \dots, i_l)$, $|I| = l$, where $1 \leq i_j \leq m$, and $l \leq k$, we define the vector field X_I by $X_I = [X_{i_1} [X_{i_2} \cdots [X_{i_{l-1}}, X_{i_l}] \cdots]$.

Next, for each n -tuple of such I 's, $\{I_1, I_2, \dots, I_n\} = J$, write $|J| = |I_1| + |I_2| + \dots + |I_n|$, and define λ_J by $\lambda_J(x) = \det(X_{I_1}, X_{I_2}, \dots, X_{I_n})(x)$. The λ_J may be thought of as generalizations of the Levi invariants, to which they reduce in special cases.

THEOREM 7. (a) *For δ sufficiently small, $\rho(x, y) < \delta$ if and only if $y = \exp(\sum_I a_I X_I)(x)$, with a_I constants satisfying $|a_I| < c\delta^{|I|}$.*

(b) $V_x(\delta) \approx \sum_J |\lambda_J(x)| \delta^{|J|}$.

From part (b) we see, in particular, that the volume of the balls have the important doubling property.

Next we state the estimates for the parametrix $K(x, y)$ of the operator $\sum_{j=1}^m X_j^2$.

THEOREM 8.

$$|X_{i_1} \cdots X_{i_r} K(x, y)| \leq c_r \frac{(\rho(x, y))^{2-r}}{V_x(\delta)} \quad (10.3)$$

where $\delta = \rho(x, y)$.³

REMARK. Notice that for homogeneous nilpotent groups the forms of all these estimates are implied by scale-invariance.

The two theorems above, and further results along these lines, are in Nagel, Stein, and Wainger [70, 71]; see also Nagel [65]. There is an alternate theory leading to Theorem 8, among other results. It is due to C. Fefferman and Phong [31], C. Fefferman [30], Sanchez-Calle [87], and C. Fefferman and Sanchez-Calle [32]. This approach has the advantage of being able to treat the general subelliptic operators which are of the form $\sum a_{ij}(x) \partial^2 / \partial x_i \partial x_j$, with $\{a_{ij}(x)\}$ semidefinite and smooth. On the other hand, the first approach described applies to the general Hörmander operator $\sum X_j^2 + X_0$ and, for that matter, to any subelliptic operator which is a "homogeneous" polynomial in the vector fields, and it gives analogues of (10.3) for their parametrics.

10. Singular Radon transforms. This topic may be viewed as the culmination in the progression from singular integrals in the translation invariant case of \mathbf{R}^n treated in §6, to the case for nilpotent groups, as in §§7 and 8, and thence to the general "variable coefficients" situation. Here the complete picture, both with respect to the optimal formulation of the general assertions and to the details of the expected proofs, is not yet fully worked out. So we will describe only an important special case which is of particular interest because of its relevance to the $\bar{\partial}$ -Neumann problem.

Let M be a (compact) manifold of dimension $m + 1$, $m > 1$. Then, imitating the situation described at the end of §7, we assume that for each $P \in M$ we are given a submanifold M_P of dimension m , with $P \in M_P$; also for each P we are to be given a Calderón-Zygmund kernel density $K(P, \cdot)$ on M_P , with singularity

³Except of course in the case where $n = 2$ and when $X_1 \cdots X_m$ already span; then (10.3) is asserted only for $r \geq 1$.

at P , so that the mappings $P \rightarrow M_P$ and $P \rightarrow K(P, \cdot)$ are smooth. We then define the singular Radon transform T (mapping $C^\infty(M)$ to itself) by

$$(Tf)(P) = \int_{M_P} K(P, \cdot) f_P(\cdot) \quad (11.1)$$

where f_P denotes the restriction of f to M_P .

The curvature condition that intervenes here turns out to be related to that occurring (in Guillemin and Sternberg [41]) in the invertibility problem for the more standard versions of the Radon transforms (e.g., when $K(P, \cdot)$ in (11.1) is everywhere smooth).

Let $\Phi(P, Q) = 0$ be a defining function for the relation $Q \in M_P$ which is nondegenerate (i.e., $d_P \Phi \neq 0$, $d_Q \Phi \neq 0$). Then the "rotational curvature" form $L_\Phi = L_\Phi(P)$ is the bilinear form defined for $(v_1, v_2) \in T_P(M_P) \times T_P(M_P)$ by $L_\Phi(v_1, v_2) = \langle d_{PQ}^2 \Phi|_{P=Q} v_1, v_2 \rangle$, where the Hessian $d_{PQ}^2 \Phi$ is the differential of the mapping $\Phi \rightarrow d_P \Phi(P, Q)$ (the differential is a linear map from $T_Q(M)$ to $T_P(M)^*$). The curvature condition is:

$$\text{The rotational curvature form } L \text{ is nondegenerate for each } P \in M. \quad (11.2)$$

THEOREM 9. *The operator T defined by (11.1) is bounded on $L^p(M)$ to itself, $1 < p < \infty$, if the curvature condition (11.2) is satisfied.*

(See Phong and Stein [77, 78]; an earlier approach for the case of variables curves in the plane is outlined in Nagel, Stein, and Wainger [69].)

REMARKS. (1) There are also L^p inequalities for a related maximal function which is associated to T in the same way as M_1 is associated to T_1 in (6.1') and (6.2').

(2) A connection between operators of the type (11.1) and Fourier integral operators with singular symbols associated to a pair of Lagrangians has been pointed out by Uhlmann [114]. For the latter class of operators see Guillemin and Uhlmann [42], also Melrose and Uhlmann [61].

(3) To clarify the curvature condition, we point out two cases: First, in the translation-invariant case in \mathbf{R}^{n+1} the condition means that M_0 has nonvanishing Gauss curvature at the origin. Second, when \mathcal{D} is a smooth domain in \mathbf{C}^{n+1} , $M = \text{boundary of } \mathcal{D}$ ($m = 2n$), then for a naturally attached family $\{M_P\}$ the condition is equivalent to the nondegeneracy of the Levi form.

The idea of the proof of the theorem is first to localize to a coordinate system, $(t, x) \in \mathbf{R} \times \mathbf{R}^m$, so that if $P = (t, x)$, $Q = (s, y)$, then the relation $Q \in M_P$ is given by $s = t + S(t, x, y)$, with $S(t, x, x) \equiv 0$, and so (11.2) becomes

$$\det \left(\frac{\partial^2 \Phi(x, y)}{\partial x_i \partial y_j} \right) \neq 0 \quad (11.3)$$

where $\Phi(x, y) = S(t, x, y)$.

After a Fourier transform in the t -variable, one can ultimately reduce the problem of the L^2 boundedness of (11.1) to oscillatory integrals of the form

$$I_\lambda: f \rightarrow \int_{\mathbf{R}^n} e^{i\lambda\Phi(x, y)} K(x, y) f(y) dy. \quad (11.4)$$

Here K is a compactly supported kernel of a standard pseudodifferential operator of order μ , with $-m < \mu \leq 0$.

LEMMA. *If Φ satisfies (11.3), and $-m < \mu \leq 0$, then for the operator norm of I_λ we can make the estimate*

$$\|I_\lambda\|_{L^2 \rightarrow L^2} \leq A(1 + |\lambda|)^{\mu/2}. \quad (11.5)$$

Note the relation between the oscillatory integral (11.4), and those in (7.5), (8.4), and (12.3) below. The special case arising when K is the kernel of an operator in $\text{Op}(S_{1,0}^{-\infty})$ is in Hörmander [50], but the order of decrease in (11.5) cannot be improved beyond $(1 + |\lambda|)^{-m/2}$. For the relation of Theorem 9 and the $\bar{\partial}$ -Neumann problem, see Phong and Stein [79].

12. *Averages over variable hypersurfaces.* We return to our first topic, maximal surface averages discussed in §§1 and 4, and try to obtain diffeomorphic-invariant versions of these results. Our setting is as follows:

For each $x \in \mathbf{R}^n$ and $\varepsilon \geq 0$, we are given a smooth hypersurface $S_{\varepsilon,x} \subset \mathbf{R}^n$. The family of surfaces $\{S_{\varepsilon,x}\}$ will be described as images of a fixed hypersurface \bar{S} : We assume that $S_{\varepsilon,x} = \rho_{\varepsilon,x}(\bar{S})$, where $\rho_{\varepsilon,x}$ is an imbedding map for each (ε, x) , and such that $\rho_{\varepsilon,x}(u)$ is smooth for $(\varepsilon, x, u) \in [0, \infty) \times \mathbf{R}^n \times \bar{S}$. Fix a cut-off function $\psi \in C_0^\infty(\mathbf{R}^n)$ and define the mean-value operator

$$M_\varepsilon(f)(x) = \int_{\bar{S}} f(x + \varepsilon \rho_{\varepsilon,x}(u)) \psi(x) d\sigma(u), \quad (12.1)$$

with $d\sigma$ a measure with smooth density and compact support on \bar{S} . The relevant curvature conditions are

- (a) The hypersurface $S_{0,x}$, for each $x \in \mathbf{R}^n$, has Gauss curvature not vanishing of infinite order at any point.
 - (b) Same as (a), except the Gauss curvature is assumed to be nowhere vanishing.
- (12.2)

Notice that these conditions are diffeomorphic invariant, in the sense that if $\Phi: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is a diffeomorphism, then

$$\Phi(x + \varepsilon \rho_{\varepsilon,x}(u)) = y + \varepsilon \tilde{\rho}_{\varepsilon,y}(u), \quad \text{if } \Phi(x) = y$$

where $\tilde{\rho}$ satisfies the same conditions ((12.2)(a) or (b)) as ρ .

THEOREM 10. *Let*

$$\mathcal{M}(f) = \sup_{0 < \varepsilon \leq \varepsilon_0} |M_{\varepsilon_0}(f)|.$$

If ε_0 is sufficiently small, the mapping $f \rightarrow \mathcal{M}(f)$ is bounded in L^p to itself in the following situations:

- (i) *when (12.2(a)) is satisfied, when $n \geq 6$, and $p > p_0$, for a finite $p_0 = p_0(\{s_{\varepsilon,x}\})$;*
- (ii) *when (12.2(b)) is satisfied, when $n \geq 3$, $p > n/(n-1)$.*

For part (ii) see Greenleaf [37], but the argument given there does not extend to cover case (i). As to the proof of that (see Sogge and Stein [93]), while

it is not, strictly speaking, modeled on nilpotent groups, it has nevertheless essential features in common with the argument of Theorem 9 for singular Radon transforms. In fact, one proceeds, after an appropriate coordinate choice and taking $n = m + 1$, to study the key square function via the Fourier transform in one variable. (It is at this stage of the proof that the restriction $n \geq 6$ enters.) One is again led to oscillatory integral operators

$$\tilde{I}_\lambda: f \rightarrow \int_{\mathbf{R}^m} e^{i\lambda\Phi(x,y)} |H(x,y)|^N \psi(x,y) f(y) dy, \quad (12.3)$$

where Φ is a real function, $H(x,y)$ is its Hessian $= \det(\partial^2 \Phi / \partial x_i \partial x_j)$, and $\psi \in C_0^\infty(\mathbf{R}^m \times \mathbf{R}^m)$. In effect, the properties of Φ and its Hessian express the curvature conditions imposed.

In analogy with Lemma 2 in §2 and the lemma in §1 one can prove the following estimate for the L^2 norm of the operators \tilde{I}_λ :

LEMMA. For N sufficiently large

$$\|\tilde{I}_\lambda\|_{L^2 \rightarrow L^2} \leq A(1 + |\lambda|)^{-m/2}. \quad (12.3)$$

In fact, (12.3) holds if $N \geq 5m/2$.

Theorem 10 suggests the following

PROBLEM. Can we extend both (i) and (ii) to $n \geq 2$?

Part IV. Some further results. We describe very briefly some related areas of research in harmonic analysis where oscillatory integrals (and curvature) play an important role.

13. *Restriction theorems and Bochner-Riesz summability.* The appellation “restriction theorem” refers to the a priori inequalities for restriction of the Fourier transform in \mathbf{R}^n to submanifolds S of the form

$$\left(\int_S |\hat{f}(\xi)|^q d\sigma(\xi) \right)^{1/q} \leq A \|f\|_{L^p}, \quad (13.1)$$

which go back to the observation (unpublished) of the author that such estimates do hold whenever the Fourier transform of $d\sigma$ has a decay at infinity like $O(|\xi|^{-\epsilon})$. Thus by Lemma 1 in §2 there are such restriction theorems for all S of finite type.

Bochner-Riesz summability is closely connected with the problem of finding L^p estimates of the Fourier multiplier operators on \mathbf{R}^n given by

$$m(\xi) = (1 - |\xi|^2)_+^\lambda, \quad (13.2)$$

and

$$S_R^\lambda(f)^\wedge = m(\xi/R) \hat{f}(\xi).$$

The results in C. Fefferman [28], Carleson and Sjölin [11], and Tomas [113] related restriction theorems for the sphere with the behavior of the multiplier (13.2), and obtained sharp L^p estimates for both problems when $n = 2$, and also partial results when $n \geq 3$. The first proofs of some of these results were

reformulated in Hörmander [50] in terms of boundedness properties of certain oscillatory integrals as mappings from $L^p(\mathbf{R}^{n-1})$ to $L^q(\mathbf{R}^n)$, with $q = (\frac{n+1}{n-1})p'$.

One considers the operators

$$(T_\lambda f)(x) = \int_{\mathbf{R}^{n-1}} e^{i\lambda\Phi(x,y)} \psi(x,y) f(y) dy, \quad x \in \mathbf{R}^n, \quad (13.3)$$

where $\psi \in C_0^\infty(\mathbf{R}^n \times \mathbf{R}^{n-1})$; for Φ one makes an assumption similar to everywhere nonvanishing Gaussian curvature.

Then one has the inequalities

$$\|T_\lambda(f)\|_{L^q(\mathbf{R}^n)} \leq A\lambda^{-n/q} \|f\|_{L^p(\mathbf{R}^{n-1})}. \quad (13.4)$$

These hold for $1 \leq p < 4$, when $n = 2$, as shown by Carleson and Sjölin, and for $n \geq 3$, when $1 \leq p \leq 2$ (see Stein [103]). The main problem left open is the range $2 < p < 2n/(n-1)$.

An alternate approach for $n = 2$ (which emphasizes the geometry of rectangles in Fourier transform space) is in C. Fefferman [29]; it was pursued by Cordoba [17] who found related L^2 estimates for rectangular maximal functions in \mathbf{R}^2 (see also §14 below). Further, Carbery [8] used this approach to prove sharp estimates for the key square function

$$\left(\int_0^\infty |S_R^\lambda - S_R^{\lambda-1}|^2 dR/R \right)^{1/2}$$

in \mathbf{R}^2 , for $2 \leq p \leq 4$, obtaining a maximal theorem for Bochner-Riesz summability. Earlier use of this square function is in, e.g., Stein and Weiss [106, Chapter 7]. For related results in this area see also Igari [53], Cordoba and López-Melero [20], Cordoba [18], Carbery [9], Christ [13], Seeger [88], and Sogge [90].

14. *Directed maximal functions; in particular, lacunary directions.* For any unit vector θ in \mathbf{R}^n define the directed maximal function M_θ by

$$M_\theta(f)(x) = \sup_{h>0} \frac{1}{2h} \left| \int_{-h}^h f(x - t\theta) dt \right|.$$

In several problems one is interested in the boundedness properties of M_Γ where $M_\Gamma(f) = \sup_{\theta \in \Gamma} M_\theta(f)$, with Γ a given set of directions. (When $\Gamma = S^{n-1}$, $n \geq 2$, there are no L^p inequalities for $p < \infty$ because of the Besicovitch-Kakeya set.)

There are two situations when M_Γ has been studied. First, when Γ is *uniformly distributed*, i.e., $\exists \varepsilon > 0$, so that if $\theta, \theta' \in \Gamma$, $\theta \neq \theta'$, then $|\theta - \theta'| \geq \varepsilon$. Then it can be conjectured that

$$\|M_\Gamma(f)\|_p \leq A(\log 1/\varepsilon)^N \|f\|_p,$$

if $p \geq n$, for some $N = N(p, n)$. This has been shown by Cordoba [17] for $n = 2$ in connection with the multiplier (13.2); see also Stromberg [111].

The second situation occurs when Γ is *lacunary*, e.g., in 2 dimensions, with $\Gamma = \{\theta_k\}$, θ_k makes an angle 2^{-k} with the x -axis. Then

$$\|M_\Gamma(f)\|_p \leq A_p \|f\|_p, \quad 1 < p < \infty. \quad (14.1)$$

This can be proved by using square functions and decay properties of the Fourier transform; see Nagel, Stein, and Wainger [68]. The argument combines a Littlewood-Paley inequality

$$\left\| \left(\sum |P_k f|^2 \right)^{1/2} \right\|_p \leq A_p \|f\|_p, \quad 1 < p < \infty,$$

where $P_k(f)^\wedge(\xi) = \chi_{S_k}(\xi) \hat{f}(\xi)$, and S_k denotes the sector

$$S_k = \{ \xi \mid |(\xi, \theta_k)| - |\xi| \mid \leq 2^{-k} |\xi| \},$$

together with a “bootstrap”: if (14.1) holds for p_0 then

$$\left\| \left(\sum |M_{\theta_k} f_k|^2 \right)^{1/2} \right\|_p \leq A \left\| \left(\sum |f_k|^2 \right)^{1/2} \right\|_p$$

holds if $1/p < \frac{1}{2}[1 + 1/p_0]$.

Earlier results for $p \geq 2$, by covering arguments, are in Stromberg [110], and Cordoba and R. Fefferman [19]. Further results relevant to M_Γ in the uniformly distributed case are in R. Fefferman [33] and Christ, Duoandikoetxea, and Rubio de Francia [15]. Bootstrap arguments similar to the one above are exploited in Duoandikoetxea and Rubio de Francia [26].

15. Hilbert transform along flat convex curves. We return to §6 and here we point out that even when the submanifold is infinitely flat at the origin, results such as Theorem 5 may still hold. We shall consider curves $\gamma(t) = (t, \gamma_2(t))$ in \mathbf{R}^2 with $t \rightarrow \gamma_2(t)$ convex, continuous, for $0 \leq t < \infty$, $\gamma_2(0) = \gamma_2'(0) = 0$, and $\gamma_2(t)$ odd (i.e., $\gamma_2(-t) = -\gamma_2(t)$).

One considers the Hilbert transform and maximal function along γ ,

$$H(f)(x) = \text{p. v.} \int_{-\infty}^{\infty} f(x - \gamma(t)) dt/t, \quad (15.1)$$

$$M(f)(x) = \sup_{h>0} \left| \frac{1}{h} \int_0^h f(x - \gamma(t)) dt \right|. \quad (15.2)$$

Then a necessary and sufficient condition for (15.1) to be bounded on $L^p(\mathbf{R}^2)$ is that the curve satisfies the “doubling-time” condition: with $h(t) = t\gamma_2'(t) - \gamma_2(t)$ there is a constant $D > 1$ so that $h(Dt) \geq 2h(t)$, $t \geq 0$. The result for $p = 2$ (as well as related results for “even” curves, some higher-dimensional variants, and (15.2) for $p = 2$) are in Nagel, Vance, Wainger, and Weinberg [72–74]. The general L^p result and the L^p inequalities for M are in Carlsson et al. [12]. The L^p argument makes use of a bootstrap which has some similarity to that in §14 above.

16. Fourier transform of surface-carried measures—a vignette. We close this survey by returning to the problem of §2 and presenting a recent elegant result of Bruna, Nagel, and Wainger [7], which provides a striking illustration of the interrelation of many of the notions considered in this essay: the decay of the Fourier transform, curvature, and quasimetrics, and their associated volume forms.

Let S denote the boundary surface of a smooth convex compact set in \mathbf{R}^n , and the question is the decay at infinity of (2.2), the Fourier transform of a smooth surface-carried measure. We assume also that S is of finite type at each point, in the stronger sense, that for each $x_0 \in S$ each tangent line passing through x_0 makes only a finite order contact with S .

Now define a family of balls $\{B(x_0, \delta)\}$, for each $x_0 \in S$, $\delta > 0$, by

$$B(x_0, \delta) = \{x \in S \mid (x - x_0, \nu_{x_0}) < \delta\},$$

where ν_{x_0} is the inward-pointing normal to S at x_0 . ($B(x_0, \delta)$ consists of those points in S which lie between the hyper-planes determined by $T_{x_0}(S)$ and $T_{x_0}(S) + \delta\nu_{x_0}$.)

Now it can be shown first that this family of balls satisfy the basic properties of the Vitali covering lemma: i.e., the quantity $\rho(x, y) = \inf\{\delta \mid y \in B(x, \delta)\}$ is a quasimetric on S , and the volume form $V_x(\delta) = \text{volume of } B(x, \delta)$, has the doubling property.

Next, for $\xi \neq 0$ let $x^+(\xi)$ and $x^-(\xi)$ denote the points on S so that

$$\nu_{x^+} = \xi/|\xi|, \quad \nu_{x^-} = -\xi/|\xi|.$$

Then the desired estimate is

$$|\hat{\mu}(\xi)| \leq c(V_{x^+}(\delta) + V_{x^-}(\delta)) \quad \text{with } \delta = 1/|\xi|. \quad (16.1)$$

REFERENCES

1. M. Beals, C. Fefferman, and R. Grossman, *Bull. Amer. Math. Soc. (N.S.)* **8** (1983), 125–322.
2. J. E. Björck, *On Fourier transforms of smooth measures carried by real-analytic sub-manifolds of \mathbf{R}^n* , Preprint, 1973.
3. J. Bourgain, *On high dimensional maximal functions associated to convex bodies*, Preprint, 1985.
4. —, *On the spherical maximal function in the plane*, Preprint, 1985.
5. —, *Averages in the plane over convex curves and maximal operators*, Preprint, 1986.
6. —, *On the L^p bounds for maximal functions associated to convex bodies in \mathbf{R}^n* , Preprint, 1986.
7. J. Bruna, A. Nagel, and S. Wainger, personal communication, 1986.
8. A. Carbery, *Duke Math. J.* **50** (1983), 409–416.
9. —, *Recent progress in Fourier analysis*, Proceedings of El Escorial Seminar, 1983, North-Holland Math. Stud. No. 111, North-Holland, Amsterdam-New York, 1985, pp. 49–56.
10. —, *Bull. Amer. Math. Soc. (N.S.)* **14** (1986), 269–273.
11. L. Carleson and P. Sjölin, *Studia Math.* **44** (1972), 287–299.
12. H. Carlsson et al., *Bull. Amer. Math. Soc. (N.S.)* **14** (1986), 263–267.
13. M. Christ, *Proc. Amer. Math. Soc.* **95** (1985), 16–20.
14. —, *Ann. of Math.* **122** (1985), 575–596.
15. M. Christ, J. Duoandikoetxea, and J. L. Rubio de Francia, *Duke Math. J.* **53** (1986), 189–209.
16. R. Coifman, A. MacIntosh, and Y. Meyer, *Ann. of Math.* **116** (1982), 361–387.
17. A. Cordoba, *Amer. J. Math.* **99** (1977), 1–22.
18. —, *Ann. Inst. Fourier (Grenoble)* **32** (1982), 215–226.
19. A. Cordoba and R. Fefferman, *Proc. Nat. Acad. Sci. U.S.A.* **74** (1977), 2211–2213.
20. A. Cordoba and B. López-Melero, *Ann. Inst. Fourier (Grenoble)* **31** (1981), 147–152.
21. M. Cowling and G. Mauceri, personal communication, 1982.

22. —, Trans. Amer. Math. Soc. **287** (1985), 431–455.
23. G. David, Proc. Internat. Congr. Math. (Berkeley, California, U.S.A., 1986), 1987.
24. G. David and J. L. Journé, Ann. of Math. **120** (1984), 371–397.
25. J. Duoandikoetxea and J. L. Rubio de Francia, personal communication, 1984.
26. —, Invent. Math. (1986).
27. K. J. Falconer, Math. Proc. Cambridge Philos. Soc. **87** (1980), 221–226.
28. C. Fefferman, Acta Math. **124** (1970), 9–36.
29. —, Israel J. Math. **15** (1973), 44–52.
30. —, Bull. Amer. Math. Soc. (N.S.) **9** (1983), 129–206.
31. C. Fefferman and D. H. Phong, Conference on Harmonic Analysis in Honor of Antoni Zygmund, 1981, Vol. 2, Wadsworth, Calif., 1983, pp. 590–606.
32. C. Fefferman and A. Sanchez-Calle, 1986 (to appear).
33. R. Fefferman, Proc. Nat. Acad. Sci. U.S.A. **80** (1983), 3877–3878.
34. G. Folland and E. M. Stein, Comm. Pure Appl. Math. **27** (1974), 429–522.
35. D. Geller and E. M. Stein, Bull. Amer. Math. Soc. (N.S.) **6** (1982), 99–103; Math. Ann. **267** (1984), 1–15.
36. R. Goodman, *Nilpotent Lie groups*, Lecture Notes in Math., Vol. 562, Springer-Verlag, Berlin and New York, 1976.
37. A. Greenleaf, Indiana Math. J. **30** (1981), 519–537.
38. —, personal communication, 1983.
39. —, *Singular integral operators with conical singularities*, Preprint, 1985.
40. A. Grossman, G. Loupias, and E. M. Stein, Ann. Inst. Fourier (Grenoble) **18** (1969), 343–368.
41. V. Guillemin and S. Sternberg, *Geometric asymptotics*, Math. Surveys, No. 14, Amer. Math. Soc., Providence, R. I. 1977.
42. V. Guillemin and G. Uhlmann, Duke Math. J. **48** (1981), 251–267.
43. R. Harvey and J. Polking, Proc. Sympos. Pure Math., Vol. 41, Amer. Math. Soc., Providence, R.I., 1985, pp. 117–136.
44. B. Helffer and J. Nourigatt, *Hypoellipticité maximale pour des opérateurs polynomes de champs de vecteurs*, Birkhauser, Basel, 1985.
45. G. M. Henkin, Math. USSR-Sb. **7** (1969), 597–616.
46. C. S. Herz, Ann. of Math. **75** (1962), 81–92.
47. E. Hlawka, Monatsh. Math. **54** (1950), 1–36.
48. L. Hörmander, Acta Math. **119** (1967), 147–171.
49. —, Acta Math. **127** (1971), 79–183.
50. —, Ark. Mat. **11** (1973), 1–11.
51. —, *The analysis of linear partial differential operators. I*, Springer-Verlag, Berlin-New York, 1983.
52. R. Howe, J. Funct. Anal. **38** (1980), 188–254.
53. S. Igari, Tôhoku Math. J. **33** (1981), 413–419.
54. A. Knapp and E. M. Stein, Ann. of Math. **93** (1971), 489–578.
55. A. Koranyi and I. Vagi, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **25** (1971), 575–648.
56. S. Krantz, *Beijing lectures on harmonic analysis*, Ann. of Math. Stud. (to appear).
57. I. Lieb, Math. Ann. **190** (1970), 6–44.
58. I. Lieb and M. Range, Math. Ann. **265** (1983), 221–251.
59. J. M. Marstrand, Mathematika **26** (1979), 180–183.
60. G. Mauceri, M. Picardello, and F. Ricci, Supp. Rend. Circ. Mat. Palermo **1** (1981), 191–202.
61. R. Melrose and G. Uhlmann, Comm. Pure Appl. Math. **32** (1979), 483–519.
62. P. A. Meyer, *Demonstrations probabiliste de certaines inegalités de Littlewood-Paley*, Lecture Notes in Math., vol. 511, Springer-Verlag, Berlin-New York, 1976, pp. 125–183; *Retour sur la théorie de Littlewood-Paley*, Lecture Notes in Math., vol. 850, Springer-Verlag, Berlin-New York, 1980, pp. 151–166.
63. D. Müller, Invent. Math. **73** (1983), 467–489.
64. —, J. Reine Angew. Math. **356** (1985), 90–118.
65. A. Nagel, *Beijing lectures on harmonic analysis*, Ann. of Math. Stud. (to appear).
66. A. Nagel, N. M. Riviere, and S. Wainger, Amer. J. Math. **98** (1976), 395–403.

67. —, Proc. Nat. Acad. Sci. U.S.A. **73** (1976), 1416–1417.
68. A. Nagel, E. M. Stein, and S. Wainger, Proc. Nat. Acad. Sci. U.S.A. **75** (1978), 1060–1062.
69. —, *Hilbert transforms and maximal functions related to variable curves*, Proc. Sympos. Pure Math., Vol. 35, Part 1, Amer. Math. Soc., Providence, R. I., 1979, pp. 95–98.
70. —, Proc. Nat. Acad. Sci. U.S.A. **78** (1981), 6596–6599.
71. —, Acta Math. **155** (1985), 103–147.
72. A. Nagel, J. Vance, S. Wainger, and D. Weinberg, Duke Math. J. **50** (1983), 735–744.
73. —, Duke Math. J. **52** (1985), 715–722.
74. —, Amer. J. Math. (to appear).
75. D. M. Oberlin and E. M. Stein, Indiana Univ. Math. J. **31** (1982), 641–650.
76. D. H. Phong, Proc. Nat. Acad. Sci. U.S.A. **76** (1979), 1554–1558.
77. D. H. Phong and E. M. Stein, Proc. Nat. Acad. Sci. U.S.A. **80** (1983), 7697–7701.
78. —, *Hilbert integrals, singular integrals, and Radon transforms*. Part I, Acta Math. **157** (1986), 99–157.
79. —, *ibid.*, Part II, Invent. Math. **86** (1986), 75–113.
80. B. Randol, Trans. Amer. Math. Soc. **139** (1969), 279–285.
81. F. Ricci and E. M. Stein, Proc. Nat. Acad. Sci. U.S.A. **83** (1986), 1–3.
82. —, *Harmonic analysis on nilpotent-groups and singular integrals*. Part I, J. Funct. Anal. (to appear).
83. —, *ibid.*, Part II, Preprint, 1986.
84. L. P. Rothschild and E. M. Stein, Acta Math. **137** (1976), 247–320.
85. J. L. Rubio de Francia, *Maximal functions and Fourier transforms*, Preprint, 1986.
86. A. Ruiz, Trans. Amer. Math. Soc. **287** (1985), 167–188.
87. A. Sanchez-Calle, Invent. Math. **78** (1984), 143–160.
88. A. Seeger, *Über Fouriermultiplikatoren und die ihnen Zugeordneten Maximal funktionen*, Thesis, Darmstadt, 1985.
89. I. Segal, Math. Scand. **13** (1963), 31–43.
90. C. Sogge, Duke Math. J. **53** (1986), 43–65.
91. —, *On almost everywhere convergence to L^p data for higher order hyperbolic operators*, Preprint, 1986.
92. C. Sogge and E. M. Stein, Invent. Math. **82** (19985), 543–556.
93. —, *Averages over hypersurfaces*. II, Invent. Math. **86** (1986), 133–242.
94. N. K. Stanton, Invent. Math. **65** (1981), 137–174.
95. E. M. Stein, Proc. Nat. Acad. Sci. U.S.A. **47** (1961), 1894–1897.
96. —, *Topics in harmonic analysis related to the Littlewood-Paley theory*, Princeton Univ. Press, Princeton, N. J., 1970.
97. —, *Singular integrals and differentiability properties of functions*, Princeton Univ. Press, Princeton, N.J., 1970.
98. —, *Some problems in harmonic analysis suggested by symmetric spaces and semi-simple groups*, Actes du Congrès International des Mathématiciens (Nice, 1970), Vol. 1, Gauthier-Villars, Paris, 1971, pp. 173–189.
99. —, Proc. Nat. Acad. Sci. U.S.A. **73** (1976), 2174–2175.
100. —, Bull. Amer. Math. Soc. (N.S.) **7** (1982), 359–376.
101. —, Bull. Amer. Math. Soc. (N.S.) **9** (1983), 71–73.
102. —, *Recent progress in Fourier analysis*, Proceedings of El Escorial Seminar, 1983, North-Holland Math. Stud. No. 111, North-Holland, Amsterdam-New York, 1985, pp. 229–244.
103. —, *Beijing lectures on harmonic analysis*, Ann. of Math. Stud. (to appear).
104. E. M. Stein and J. O. Stromberg, Ark. Mat. **21** (1983), 259–269.
105. E. M. Stein and S. Wainger, Bull. Amer. Math. Soc. **84** (1978), 1239–1295.
106. E. M. Stein and G. Weiss, *Introduction to Fourier analysis on Euclidean spaces*, Princeton Univ. Press, Princeton, N.J., 1971.
107. R. S. Strichartz, Duke Math. J. **44** (1977), 705–713.
108. —, Duke Math. J. **48** (1981), 699–727.
109. —, Studia Math. **74** (1982), 137–151.
110. J. O. Stromberg, Ark. Mat. **15** (1977), 229–240.
111. —, Ann. of Math. **107** (1978), 399–402.

112. I. Svenson, Ark. Mat. **9** (1971), 11–22.
113. P. Tomas, Proc. Sympos. Pure Math., Vol. 35, Part 1, Amer. Math. Soc., Providence, R. I., 1979, pp. 111–114.
114. G. Uhlmann, personal communication, 1986.
115. A. Zygmund, *Trigonometric series*, Cambridge Univ. Press, London and New York, 1959.

PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08544, USA

Algebraic K -Theory of Fields

A. A. SUSLIN

The construction of higher algebraic K -theory was achieved by the fundamental work of Quillen [24]. After that the main efforts were concentrated in the field of computations and applications of K -theory to concrete algebraic problems. The most intriguing are the conjectures relating algebraic K -theory to étale cohomology. Such conjectures in certain particular cases were made by Quillen and Lichtenbaum [14, 9, 23]. Nowadays all conjectures of this type are usually called the Quillen-Lichtenbaum conjectures. One of the important properties of algebraic K -theory is the exact localization sequence: if $Y \subset X$ is a closed subscheme, then there is a long exact sequence

$$\cdots \rightarrow K'_i(Y) \rightarrow K'_i(X) \rightarrow K'_i(X - Y) \xrightarrow{\partial} K'_{i-1}(Y) \rightarrow \cdots$$

($K' = K$ for regular schemes) and the resulting spectral sequence

$$E_1^{pq} = \coprod_{\text{codim } x=p} K_{-p-q}(k(x)) \Rightarrow K'_{-p-q}(X).$$

Another important property is Gersten's conjecture, proved by Quillen, which makes it possible to identify the second term of this spectral sequence: $E_2^{pq} = H^p(X, K_{-q})$. These properties often reduce general problems of algebraic K -theory to the particular case of fields in which case these problems are especially explicit and intriguing.

Higher K -theory of a field F (as well as of any ring) may be defined in terms of Quillen's plus construction: $K_i(F) = \pi_i(\text{BGL}(F)^+)$, where $\text{BGL}(F)^+$ is the H -space having the same homology as $\text{BGL}(F)$, i.e., the same as homology of the discrete group $\text{GL}(F)$. Thus K -theory is closely related to the homology theory of $\text{GL}(F)$.

This paper concerns some of the recent achievements in the K -theory of fields and in related areas. To a pity, I have only mentioned very briefly such an important field as étale K -theory of Dwyer-Friedlander; the ideas and methods used in this theory are very far from those discussed in the main part of this paper.

1. Norm-residue homomorphism. In view of the Moore-Matsumoto theorem, the group $K_2(F)$ may be described as a group with generators $\{a, b\}$ ($a, b \in F^*$) and relations $\{a_1 a_2, b\} = \{a_1, b\} + \{a_2, b\}$, $\{a, b_1 b_2\} = \{a, b_1\} + \{a, b_2\}$, $\{a, 1 - a\} = 0$ ($a \neq 1$). Suppose that n is an integer prime to $\text{char } F$; then we have the Kummer isomorphism $\chi: F^*/F^{*n} \xrightarrow{\sim} H^1(F, \mu_n)$. It is easy to verify that $\chi(a) \cup \chi(1-a) = 0 \in H^2(F, \mu_n^{\otimes 2})$ and hence we get a well-defined homomorphism

$$R_n = R_{n,F}: K_2(F)/n \rightarrow H^2(F, \mu_n^{\otimes 2}): \{a, b\} \mapsto \chi(a) \cup \chi(b),$$

which is called the norm-residue homomorphism. In case $F \supset \mu_n$ the choice of the primitive n th root of unity ξ makes it possible to identify G_F -modules μ_n and $\mu_n^{\otimes 2}$ and hence to identify $H^2(F, \mu_n^{\otimes 2})$ with $H^2(F, \mu_n) = {}_n \text{Br}(F)$. After this identification R_n turns into a cyclic algebra homomorphism: $\{a, b\} \mapsto [A_\xi(a, b)]$ (cf. [19]). Thus in this case the question about surjectivity of R_n is equivalent to the classical problem of Albert whether every algebra of exponent n is similar to a product of cyclic algebras.

THEOREM 1.1 [17, 30]. *For any field F and any n prime to $\text{char } F$, $R_n: K_2(F)/n \rightarrow H^2(F, \mu_n^{\otimes 2})$ is an isomorphism.*

The general case of the theorem may be easily reduced to the case (which we will consider below) when $n = p$ is prime and $F \supset \mu_p$. There are two different, but closely related, approaches to the proof of (1.1). Both approaches use essentially the computation of certain K -cohomology groups of Severi-Brauer varieties.

The first method, the original method of Merkurjev [16], works mostly for $p = 2$. Set provisionally $k_2 = K_2/2$. Suppose that $E = F(\sqrt{a})$ is a quadratic extension of a field F and denote by $\chi(a) \in H^1(F, \mu_2)$ the cohomology class corresponding to a under the Kummer isomorphism. The exact cohomology sequence

$$H^1(F, \mu_2) \xrightarrow{\chi(a)} H^2(F, \mu_2) \rightarrow H^2(E, \mu_2) \xrightarrow{N_{E/F}} H^2(F, \mu_2)$$

shows that the validity of (1.1) implies the exactness of the sequence

$$F^*/F^{*2} \xrightarrow{\alpha} k_2(F) \rightarrow k_2(E) \xrightarrow{N_{E/F}} k_2(F). \quad (1.1.1)$$

Vice versa, if (1.1.1) is exact for any quadratic extension, then an easy inductive argument proves (1.1). Moreover, it is shown in [16] that even the exactness of

$$k_2(F) \rightarrow k_2(E) \rightarrow k_2(F)$$

for any E/F is sufficient to finish the proof. Every element of $k_2(E)$ may be written in the form $\sum_{i=1}^n \{x_i + \sqrt{a}y_i, z_i\}$ with $x_i, y_i, z_i \in F$. The norm of this element in $k_2(F)$ is equal to $\sum_{i=1}^n \{x_i^2 - ay_i^2, z_i\}$. It is not difficult to write down explicitly when the last element is equal to zero: this is equivalent (after certain cosmetic changes, including possible enlargement of n) to the existence

of certain elements u_S, v_S for every nonempty $S \subset \{1, \dots, n\}$ such that, setting $z_S = \prod_{i \in S} z_i$ we will have formulae

$$x_i^2 - y_i^2 a = \prod_{S \ni i} (u_S^2 - z_S v_S^2). \quad (1.1.2)$$

Denote by F_0 the prime subfield of F and set $F_1 = F_0(a)$. Equations (1.1.2) define an affine variety T over F_1 , and elements x_i, y_i, z_i, u_S, v_S define an F -valued point of this variety. Denoting the corresponding coordinate functions by $X_i, Y_i, Z_i, U_S, V_S \in F_1(T)$, we get in $k_2(F_1(T)(\sqrt{a}))$ the "universal" element with trivial norm $\sum_{i=1}^n \{X_i + \sqrt{a}Y_i, Z_i\}$. It is sufficient to show that this universal element lies in $k_2(F_1(T))$ —the specialization argument finishes the proof. To prove the last statement it is sufficient to show that R_2 is an isomorphism for $F_1(T)$ and $F_1(T)(\sqrt{a})$. This is trivial for the second field since this field is purely transcendental over $F_1(\sqrt{a})$ (both kernel and cokernel of R_n do not change under purely transcendental extensions [4]). The field $F_2 = F_1(Z_i, U_S, V_S)$ is purely transcendental over F_1 , and $F_1(T)$ is obtained from F_2 passing several times to the function field on a conic, given by an equation of the form $X^2 - Y^2 a = *$. So the theorem follows from

PROPOSITION 1.2 [27]. *Suppose that $\text{char } k \neq 2$, $a, b \in k^*$, and denote by F the function field on the conic, given by equation $X^2 - aY^2 = b$. If $R_{2,k}$ and $R_{2,k(\sqrt{a})}$ are isomorphisms, then $R_{2,F}$ is also an isomorphism.*

The proof of (1.2) is based on the computation of certain K -cohomology groups of the conic; it also uses extensively the theory of quadratic forms, which does not allow the use of this method for $p \neq 2$.

The second approach to the proof of (1.1), developed in [17, 28, 30], is in a certain sense opposite to the one discussed above. The main technical result in this approach is

PROPOSITION 1.3. *Suppose that $p \neq \text{char } F$ and F contains a primitive p th root of unity ξ . Let $a, b \in F^*$ and denote by X the Severi-Brauer variety, corresponding to the cyclic algebra $D = A_\xi(a, b)$. The natural homomorphisms $\ker R_{p,F} \rightarrow \ker R_{p,F(X)}$ and $\text{coker } R_{p,F} \rightarrow \text{coker } R_{p,F(X)}$ are injective.*

Assuming (1.3), one can finish the proof of (1.1) as follows. It is well known that the maps $\ker R_{p,F} \rightarrow \ker R_{p,E}$, $\text{coker } R_{p,F} \rightarrow \text{coker } R_{p,E}$ are injective if E is algebraic over F of degree prime to p , so, using (1.3), one can construct an extension \tilde{F}/F such that

- (a) all cyclic p -algebras over \tilde{F} are trivial,
- (b) \tilde{F} has no extensions of degree prime to p , and
- (c) $\ker R_{p,F} \hookrightarrow \ker R_{p,\tilde{F}}$, $\text{coker } R_{p,F} \hookrightarrow \text{coker } R_{p,\tilde{F}}$.

A classical result of Milnor [19] shows that (a) is equivalent to the equality $K_2(\tilde{F})/p = 0$. Hence $\text{Ker } R_{p,\tilde{F}} = 0$. Moreover, it is easy to see that (a) and (b) imply that $\text{Br}(\tilde{F}) = 0$ and hence $\text{coker } R_{p,\tilde{F}} = 0$. Now property (c) shows that $\ker R_{p,F} = \text{coker } R_{p,F} = 0$.

The proof of (1.3), as well as the proof of (1.2), is based on the computation of K -cohomology groups of Severi-Brauer varieties.

PROPOSITION 1.4. *In conditions of (1.3), $H^1(X, K_2) = N = \text{Nrd } D^* \subset F^*$, the natural map $K_2(F) \rightarrow H^0(X, K_2)$ is surjective.*

To prove this, one has to consider the spectral sequence $E_2^{i,j} = H^i(X, K_{-j}) \Rightarrow K_{-i-j}(X)$. The theory of Chern classes and the Riemann-Roch theorem makes it possible to show that all differentials in this spectral sequence starting at or coming to $E^{i,j}$ terms with $i + j = 0, -1$ are killed by $(\dim X)!$. In our case $\dim X = p - 1$ and we know also that all differentials in this spectral sequence are killed by p (since D has splitting fields of degree p over F). This shows that there are no differentials starting at or coming to $E^{i,j}$ with $i + j = 0, -1$ and hence $H^1(X, K_2) = E_2^{1,-2} = E_\infty^{1,-2} = K_1(X)^{1/2}$. K -theory of Severi-Brauer varieties was computed by Quillen [24]:

$$K_i(X) = K_i(F) \oplus K_i(D) \oplus \cdots \oplus K_i(D^{\otimes(p-1)}).$$

Thus to finish the proof of the first statement it is sufficient to compute the topological filtration on $K_1(X) = F \oplus N \oplus \cdots \oplus N$, which is not difficult to do. Vanishing of all differentials starting at $E^{0,-2}$ imply that the edge homomorphism

$$K_2(F) \oplus K_2(D) \oplus \cdots \oplus K_2(D^{\otimes(p-1)}) = K_2(X) \rightarrow H^0(X, K_2) = E_2^{0,-2}$$

is surjective. To finish the proof of the second statement we have to show that the image of $K_2(D^{\otimes i})$ in $H^0(X, K_2) \subset K_2(F(X))$ is contained in the image of $K_2(F)$. This requires additional information about K_2 for algebras of prime index; see (3.1) below.

Proposition 1.4 is not yet sufficient for the proof of (1.3); one needs a more precise statement that $K_2(F) = H^0(X, K_2)$. This requires information about torsion in $K_2(F)$. The basic result in this direction is Hilbert's Theorem 90 for K_2 .

THEOREM 1.5. *Let E/F be a cyclic extension of prime degree p and let σ be a generator of $\text{Gal}(E/F)$. The following sequence is exact:*

$$K_2(E) \xrightarrow{1-\sigma} K_2(E) \xrightarrow{N_{E/F}} K_2(F). \quad (1.5.1)$$

The exactness of (1.5.1) is easily proved provided the norm map $N: E^* \rightarrow F^*$ is surjective—in this case, one constructs explicitly the homomorphism $K_2(F) \rightarrow K_2(E)/(1-\sigma)K_2(E)$ inverse to $N_{E/F}$ by means of the formula $\{a, b\} \mapsto \{\alpha, b\} \bmod (1-\sigma)K_2(E)$, where $N(\alpha) = a$. Using the same trick as above we see now that we will be done if we are able to prove that if X is a Severi-Brauer variety, corresponding to a cyclic algebra $(E/F, \sigma, a)$, then the map

$$\ker N_{E/F}/(1-\sigma)K_2(E) \rightarrow \ker N_{E(X)/F(X)}/(1-\sigma)K_2(E(X))$$

is injective. This problem is simplified by the fact that the algebra under consideration splits over E and hence $X_E = \mathbf{P}_E^{p-1}$. The proof uses, in fact, only the computation of $H^1(X, K_2)$ fulfilled above.

Applying (1.5) to the universal Kummer extension $F(\sqrt[n]{T})/F(T)$ (T is transcendental over F) or to the universal Artin-Schreier extension we get the following result, which was conjectured by Tate [37].

COROLLARY 1.6. *If F contains a primitive n th root of unity ξ , then ${}_nK_2(F) = \{\xi, F^*\}$, $K_2(F)$ does not have p -torsion with $p = \text{char } F$.*

To finish the description of torsion in K_2 , one needs the description of those elements $x \in F^*$ for which $\{\xi, x\} = 0$. This question is settled by

THEOREM 1.7 [28, 30]. *Suppose that F contains a primitive n th root of unity ξ and denote by F_0 the subfield of constants in F (i.e., the algebraic closure of the prime subfield). For $x \in F^*$ the following conditions are equivalent:*

- (a) $\{\xi, x\} = 0 \in K_2(F)$;
- (b) $x = x_0 y^n$, where $y \in F^*$, $x_0 \in F_0^*$, and $\{\xi, x_0\} = 0 \in K_2(F_0)$.

COROLLARY 1.8. *If E/F is an extension such that F is algebraically closed in E , then $K_2(F) \hookrightarrow K_2(E)$.*

The last corollary shows that in conditions of (1.4), $K_2(F) \xrightarrow{\sim} H^0(X, K_2)$. Using this and (1.4), one easily finishes the proof of (1.3); see [17, 30].

The proof of (1.7) is based on the study of certain l -adic cohomology groups. The crucial role plays the following fact, related to Weil's theorem about eigenvalues of Frobenius substitution on a Tate module of an abelian variety.

PROPOSITION 1.9. *Suppose that F is finitely generated and $l \neq \text{char } F$. Then $H^1(F_0, Z_l(2)) \xrightarrow{\sim} H^1(F, Z_l(2))$ and $H^2(F_0, Z_l(2)) \hookrightarrow H^2(F, Z_l(2))$.*

The above results have many important applications in algebra and algebraic geometry, some of which may be found in [30, 39, 40, 41]. We will only mention the following, for further use.

PROPOSITION 1.10 [28, 30]. *If X/F is a complete rational variety, then $H^0(X, K_2) = K_2(F)$.*

2. Algebraically closed and local fields. Since the étale cohomology groups of an algebraically closed field are trivial, it is reasonable to expect that K -groups of such a field will also have a sufficiently simple structure. The following is one of the Quillen-Lichtenbaum conjectures (see [9, 23]).

(2.1) If F is an algebraically closed field, then $K_i(F)$ is divisible for $i \geq 1$, the torsion subgroup in $K_i(F)$ being zero if i is even, and isomorphic to $\prod_{l \neq \text{char } F} Q_l/Z_l(n)$ if $i = 2n - 1$.

This conjecture is clearly true for $i = 1$ and may be easily proved for $i = 2$ [2]. Apart from these trivial cases, the conjecture was known to be true in the case where F is the algebraic closure of a finite field [22]. For fields of positive

characteristics, this conjecture was proved in [31]. The basic result of [31] is

THEOREM 2.2. *If F/F_0 is an extension of algebraically closed fields, then for any integer n the induced maps $K_i(F_0)/n \rightarrow K_i(F)/n$, ${}_nK_i(F_0) \rightarrow {}_nK_i(F)$, $K_i(F_0, Z/n) \rightarrow K_i(F, Z/n)$ are bijective.*

This theorem is a particular case of a certain simple general principle. Let V be a contravariant functor on an appropriate category of schemes with values in the category of torsion abelian groups. Suppose further that for any finite flat morphism $X \rightarrow Y$ we are given a transfer homomorphism $N_{X/Y}: V(X) \rightarrow V(Y)$, satisfying the usual properties. Suppose finally that V is homotopy invariant, i.e., $V(X \times A^1) = V(X)$ for any X .

PROPOSITION 2.3 (RIGIDITY THEOREM). *Let X/F be a connected variety over an algebraically closed field. Then for any two points $x, y: \text{Spec}(F) \rightarrow X$, the induced maps $V(X) \rightrightarrows V(\text{Spec } F) = V(F)$ coincide.*

It is clearly sufficient to treat the case of a smooth affine curve. Consider the bilinear pairing $\text{Div}(X) \times V(X) \rightarrow V(F)$ given by $x \times u \mapsto x^*(u)$. We have to show that its restriction on $\text{Div}^0(X) \times V(X)$ is trivial. Denote by \bar{X} the smooth projective model of X and set $X_\infty = \bar{X} - X$. If f is a rational function on \bar{X} , defined and equal to one on X_∞ , then the principal divisor (f) lies in the kernel of our pairing: f defines a covering $X_0 \rightarrow \mathbf{A}_F^1 = \mathbf{P}_F^1 - 1$, where X_0 is obtained from X by deleting points where f is equal to one. The usual properties of transfer imply that the image of $(f) \times u$ in $V(F)$ coincides with the image of $(0 - \infty) \times N_{X_0/\mathbf{A}_F^1}(u|_{X_0})$, which is zero in view of homotopy invariance. Thus our pairing factors through $\text{Pic}^0(\bar{X}, X_\infty) \otimes V(X)$. The group $\text{Pic}^0(\bar{X}, X_\infty)$ coincides with the group of F -points of the corresponding Rosenlicht jacobian of \bar{X} (see [26]) and hence is divisible. Since $V(X)$ is torsion we deduce that $\text{Pic}^0(\bar{X}, X_\infty) \otimes V(X) = 0$.

COROLLARY 2.3.1. *Let F/F_0 be an extension of algebraically closed fields and let X_0/F_0 be a connected variety. If $x, y: \text{Spec } F \rightarrow X_0$ are any two F_0 -points, then the induced maps $V(X_0) \rightrightarrows V(F)$ coincide.*

COROLLARY 2.3.2. *In conditions of (2.3.1) for any F_0 -point $x: \text{Spec } F \rightarrow X_0$, the image of the corresponding homomorphism $V(X_0) \rightarrow V(F)$ is contained in the image of $V(F_0)$.*

Choose a rational point $\text{Spec } F_0 \rightarrow X_0$ and apply (2.3.1) to x and $y: \text{Spec } F \rightarrow \text{Spec } F_0 \rightarrow X_0$.

COROLLARY 2.3.3. *Suppose, in addition, that V commutes with limits:*

$$V(\text{Spec } \varinjlim A_i) = \varinjlim V(\text{Spec } A_i).$$

Then $V(F) = V(F_0)$ for any extension F/F_0 of algebraically closed fields.

F may be written as $\varinjlim A$ where A runs through all finitely generated F_0 -subalgebras of F . Our conditions imply that $V(F) = \varinjlim V(\text{Spec } A)$. The homomorphism $V(\text{Spec } A) \rightarrow V(F)$ is induced by a F_0 -point $\text{Spec } F \rightarrow \text{Spec } A$, corresponding to the imbedding $A \hookrightarrow F$. In view of (2.3.2) the image of this homomorphism is contained in the image of $V(F_0)$. Since this is true for any A we deduce that $V(F_0) \rightarrow V(F)$ is surjective. The injectivity of this map is trivial.

The present proof of (2.3), which is a slight modification of the original proof of the author [31], is due to Gabber, Gillet, and Thomason. The use of relative Picard groups instead of absolute ones allowed these authors to prove the following important generalization of (2.3).

PROPOSITION 2.4. *Let O be a henselian ring with field of fractions F and residue field k and let $X/\text{Spec } O$ be a smooth affine curve. Further, let $x, y: \text{Spec } O \rightarrow X$ be two sections that coincide in the closed point of $\text{Spec } O$. Suppose, in addition, that*

- (a) $nV(X) = 0$, where $(n, \text{char } k) = 1$,
- (b) $V(O) \hookrightarrow V(F)$.

Then the induced maps $x^, y^*: V(X) \rightarrow V(O)$ coincide.*

Choose a projective closure \bar{X} of X and set $X_\infty = \bar{X} - X$. The sections x, y define relative divisors D_x, D_y on \bar{X} (relative to X_∞). Their difference is divisible by n in $\text{Pic}(\bar{X}, X_\infty)$ since $\text{Pic}(\bar{X}, X_\infty)/n \hookrightarrow H_{\text{et}}^2(X, j_!(\mu_n))$ (where $j: X \hookrightarrow \bar{X}$) and $H_{\text{et}}^2(X, j_!(\mu_n)) = H_{\text{et}}^2(X_0, (j_0)_!(\mu_n))$, where X_0 is the closed fiber of X , in view of the proper base change theorem in étale cohomology. In view of condition (b) it is sufficient to prove the coincidence of maps $V(X_F) \rightrightarrows V(F)$, where X_F is the generic fiber of X . The last fact follows in the same manner as in the proof of (2.3), since the difference of the corresponding points is divisible by n in the relative Picard group of X_F .

REMARK 2.5. Condition (b) of Proposition 2.4 is often satisfied in algebraic K -theory in view of Quillen's theorem [24]. Moreover, this condition may be avoided in many cases of interest.

Using induction and tricks similar to those used in the proof of (2.3.3) we deduce from (2.4) the following important

THEOREM 2.6 (GABBER (UNPUBLISHED), GILLET AND THOMASON [10]). *Let V/F be a smooth variety and let $v \in V$ be a rational point. Denote by O_v^h the henselization of a local ring O_v . For any m prime to $\text{char } F$, the natural homomorphism $K_*(O_v^h, \mathbb{Z}/m) \rightarrow K_*(F, \mathbb{Z}/m)$ is bijective.*

Denote by I_v^h the maximal ideal of the local ring O_v^h . It is not difficult to deduce from (2.6) that $H_i(\text{GL}(O_v^h/I_v^h), \mathbb{Z}/m) = 0$ ($i \geq 1$); see [32]. Consider now the simplicial scheme BGL_n/F and denote by $X_{n,i}^h$ the henselization of $(\text{BGL}_n)_i = (\text{GL}_n)^i$ in unity; denote further by $O_{n,i}^h$ the coordinate ring of $X_{n,i}^h$ and by $I_{n,i}^h$ its maximal ideal. Since face and degeneracy maps of BGL_n respect

unity, we see that $X_{n,i}^h$ also form a simplicial scheme. The evident maps $X_{n,i}^h \rightarrow \mathrm{GL}_n^i \xrightarrow{\mathrm{pr}_*} \mathrm{GL}_n$ define matrices $a_k \in \mathrm{GL}_n(O_{n,i}^h, I_{n,i}^h)$. We will denote by $u_{n,i}$ the chain $[a_1, \dots, a_i] \in C_i(\mathrm{GL}_n(O_{n,i}^h, I_{n,i}^h), Z/m)$. Now one constructs, using induction on i and the fact that $H_i(\mathrm{GL}(O_{n,i}^h, I_{n,i}^h), Z/m) = 0$, chains $c_{n,i} \in C_{i+1}(\mathrm{GL}(O_{n,i}^h, I_{n,i}^h), Z/m)$ such that

$$d(c_{n,i}) = u_{n,i} - \sum_{j=0}^i (-1)^j (d_j)^*(c_{n,i-1}).$$

These considerations enable one to generalize (2.6):

COROLLARY 2.7. *Let (R, I) be a henselian pair, where R is an F -algebra. Then $\tilde{H}_*(\mathrm{GL}(R, I), Z/m) = 0$ and $K_*(R, Z/m) \xrightarrow{\sim} K_*(R/I, Z/m)$.*

The second statement follows from the first one. For the proof of the first statement it is sufficient to show that the imbedding $\tilde{C}_*(\mathrm{GL}_n(R, I), Z/m) \hookrightarrow \tilde{C}_*(\mathrm{GL}(R, I), Z/m)$ is nul-homotopic. Consider matrices $b_1, \dots, b_i \in \mathrm{GL}_n(R, I)$. These matrices define a morphism $\mathrm{Spec} R \rightarrow \mathrm{GL}_n^i$, taking $\mathrm{Spec} R/I$ to unity. Since (R, I) is a henselian pair, this morphism factors uniquely through a morphism $f_b: \mathrm{Spec} R \rightarrow X_{n,i}^h$. The desired nul-homotopy may be defined now by a formula $s([b_1, \dots, b_i]) = (f_b)^*(c_{n,i})$.

The same method of evaluation of "universal homotopy operators" $c_{n,i}$ may be applied also in many other situations. The following results are proved in [32].

THEOREM 2.8. *Let R be a henselian discrete valuation ring with maximal ideal I , fraction field F , and residue field k . For any m prime to $\mathrm{char} F$ we have canonical isomorphisms of pro-groups:*

$$\begin{aligned} H_*(\mathrm{GL}(R), Z/m) &\rightarrow \{H_*(\mathrm{GL}(R/I^n), Z/m)\}_n, \\ K_*(R, Z/m) &\rightarrow \{K_*(R/I^n), Z/m\}_n. \end{aligned}$$

To deduce the second statement from the first one it is necessary to use a version of Hurewicz's theorem for pro-spaces, proved by Panin [43].

COROLLARY 2.8.1. *In conditions of Theorem 2.8,*

$$K_*(R, Z/m) \xrightarrow{\sim} K_*(k, Z/m)$$

provided that $(m, \mathrm{char} k) = 1$.

COROLLARY 2.8.2. *Let k be an algebraically closed field of positive characteristics p and let F be the algebraic closure of the fraction field of the ring of Witt vectors over k . Then for any m prime to p there are canonical isomorphisms $K_*(k, Z/m) = K_*(F, Z/m)$.*

This corollary together with Theorem 2.2 shows that the groups $K_i(F, Z/m)$ do not depend on the algebraically closed field F (provided that m is prime to $\mathrm{char} F$); this enables us to finish the proof of the Quillen-Lichtenbaum conjecture for fields of zero characteristics. It is more natural, however, to apply the method of universal homotopy operators to the proof of the following theorem.

THEOREM 2.9 [32]. *Let F denote either the field \mathbf{R} of real numbers or the field \mathbf{C} of complex numbers. The natural morphism $\mathrm{BGL}(F)^+ \rightarrow \mathrm{BGL}(F)^{\mathrm{top}}$ induces isomorphisms on homology and homotopy groups with finite coefficients.*

Using, in addition, the Stability Theorem 4.6 (see below) we get the following result, confirming partially the isomorphism conjecture of Friedlander-Milnor [21].

COROLLARY 2.9.1. $\mathrm{BGL}_n(F) \rightarrow \mathrm{BGL}_n(F)^{\mathrm{top}}$ induce isomorphisms on $H_i(-, \mathbb{Z}/m)$ with $i \leq n$.

COROLLARY 2.9.2. *Modulo uniquely divisible groups, the K -theory of the fields \mathbf{R} and \mathbf{C} is as displayed in the following table.*

$i \bmod 8$	0	1	2	3	4	5	6	7
$K_i(\mathbf{R})$	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Q}/\mathbb{Z}	0	0	0	\mathbb{Q}/\mathbb{Z}
\downarrow	0	incl.	0	2	0	0	0	iso
$K_i(\mathbf{C})$	0	\mathbb{Q}/\mathbb{Z}	0	\mathbb{Q}/\mathbb{Z}	0	\mathbb{Q}/\mathbb{Z}	0	\mathbb{Q}/\mathbb{Z}

REMARK 2.10. A different and more algebraic approach to the proof of the Quillen-Lichtenbaum conjecture was proposed later by Jardine [42]. His method is also based on the use of (2.6).

3. The K -theory of division algebras. The K_2 -theory of division algebras was already used above in the proof of (1.4). The result that was necessary there looks as follows. If D/F is a central simple algebra and E/F is its splitting field of finite degree, then we can consider the canonical homomorphism $g_E: K_2(E) = K_2(D_E) \xrightarrow{N_{E/F}} K_2(D)$.

THEOREM 3.1 [17]. *If index of algebra D is squarefree, then $K_2(D)$ is generated by images of $K_2(E)$ over all finite splitting fields E/F .*

REMARK 3.1.1. It seems possible that the restriction on index is not really necessary for the validity of (3.1). This is a rather interesting problem. The same question may be asked for higher K -groups.

Let X/F be the Severi-Brauer variety corresponding to D . In view of (1.10) $H^0(X, K_2) = K_2(F)$ and we get a canonical homomorphism $\mathrm{Nrd}: K_2(D) \rightarrow K_2(X) \rightarrow H^0(X, K_2) = K_2(F)$.

THEOREM 3.2 [17]. *Let D be an algebra of squarefree degree over a field F .*

(a) *If $c.d.F = 2$, then $\mathrm{Nrd}: K_2(D) \xrightarrow{\sim} K_2(F)$.*

(b) *If F is a global field, then there is an exact sequence $0 \rightarrow K_2(D) \xrightarrow{\mathrm{Nrd}} K_2(F) \rightarrow \coprod_v \mathbb{Z}/2 \rightarrow 0$ where v runs through real points of F in which D is nontrivial.*

Theorem 3.2 follows easily from (3.1) and Hilbert's Theorem 90 for K_2 . I am sure that the assumption about index is superfluous for its validity. The following important result of Merkurjev [18] is much deeper.

THEOREM 3.3. *For any quaternion algebra D/F , the reduced norm $\text{Nrd}: K_2(D) \rightarrow K_2(F)$ is injective.*¹

For the proof of (3.3) Merkurjev uses his method of universal problems (cf. §1). In this case the field of definition of the universal problem is the function field on the product of three-dimensional quadrics. To prove the analog of (1.2) it is necessary to compute certain D -operator K -cohomology groups of these quadrics. The first step in this direction is provided by the theorem of Swan [36], which computes the operator K -theory of an arbitrary quadric. Further, one has to study the differentials in the BGQ-spectral sequence. The theory of Chern classes does not help this time, but Merkurjev has invented a direct method to prove the vanishing of the necessary differentials.

REMARK 3.4. (a) Merkurjev has given also the description of the image of reduced norm.

(b) It is reasonable to expect that injectivity of reduced norm holds for any algebra of squarefree degree, but at present I do not see how to attack this problem.

In the local case, to compute the K -theory of a division algebra one can use a version of methods of the previous section.

THEOREM 3.4 [35]. *Let R be a henselian discrete valuation ring with fraction field F and let D be a division algebra over F . Denote by A the maximal order in D and by I its maximal ideal. For any m prime to $\text{char } F$ there is a canonical isomorphism of pro-groups $K_i(A, Z/m) \rightarrow \{K_i(A/I^n, Z/m)\}_n$.*

COROLLARY 3.4.1. *If m is prime to $\text{char } \overline{R}$, then*

$$K_i(A, Z/m) \xrightarrow{\sim} K_i(A/I, Z/m).$$

COROLLARY 3.4.2. *Let F be a usual local field (i.e., a finite extension of the field of p -adic numbers) and D/F a division algebra of degree prime to p . For all $i \geq 1$ there are canonical isomorphisms $\text{Nrd}: K_i(D) \xrightarrow{\sim} K_i(F)$.*

COROLLARY 3.4.3 [11]. *In conditions of (3.4.2), $\text{Nrd}: K_2(D) \xrightarrow{\sim} K_2(F)$ for any division algebra D .*

Acting as in the proof of (2.9) we get, moreover,

PROPOSITION 3.5. *Denote by H the classical quaternion algebra over \mathbf{R} . The natural map $\text{BGL}(H)^+ \rightarrow \text{BGL}(H)^{\text{top}}$ induce isomorphisms on homology and homotopy groups with finite coefficients.*

4. Milnor K -theory. For any field F , its Milnor ring $K_*^M(F)$ is defined as a quotient ring of the tensor algebra $T(F^*)$ by a homogeneous ideal, generated by tensors $a \otimes (1 - a) \in T_2(F^*) = F^* \otimes F^*$ [2]. The image of $a_1 \otimes \cdots \otimes a_n$ in $K_n^M(F)$ will be denoted $\{a_1, \dots, a_n\}$. There is a canonical ring homomorphism $K_*^M(F) \rightarrow K_*(F)$, which is isomorphic in degrees ≤ 2 . The example of finite

¹Remark added in proof: Theorem 3.3 was proved independently by M. Rost [47].

fields shows that in degrees $n \geq 3$ the map $K_n^M(F) \rightarrow K_n(F)$ is not in general surjective; however, I know of no examples where this map is not injective (cf. (4.7.1) below).² The following conjecture is principal in the understanding of the structure of $K_*(F)$.

CONJECTURE 4.1. Denote by F_0 the subfield of constants in F . The ring $K_*(F)$ is generated by $K_1(F) = F^*$ and $K_*(F_0)$.³

In positive characteristics, (4.1) would imply that the homomorphism $K_*^M(F) \rightarrow K_*(F)$ is an isomorphism modulo torsion.

The following conjecture is a particular case of the general conjectures of Beilinson (see §7).

CONJECTURE 4.2. For any field F and any n prime to $\text{char } F$, the norm-residue homomorphism $K_*^M(F)/n \rightarrow \prod_{i \geq 0} H^i(F, \mu_n^{\otimes i}): \{a_1, \dots, a_i\} \mapsto \chi_n(a_1) \cup \dots \cup \chi_n(a_i)$ is an isomorphism of rings.

Another interesting conjecture concerning $K_*^M(F)$ is Milnor's conjecture about quadratic forms [20]. Suppose that $\text{char } F \neq 2$ and denote by $W(F)$ the Witt ring of nondegenerate quadratic forms over F . Let $I(F)$ denote the maximal ideal of $W(F)$, consisting of even-dimensional forms. For any $a \in F^*$ set $\langle\langle a \rangle\rangle = 1 \perp -a \in I(F)$ and $\langle\langle a_1, \dots, a_n \rangle\rangle = \langle\langle a_1 \rangle\rangle \cdot \dots \cdot \langle\langle a_n \rangle\rangle$. $I(F)$ is additively generated by $\langle\langle a \rangle\rangle$ and hence $I^n(F)$ is additively generated by n -fold Pfister forms $\langle\langle a_1, \dots, a_n \rangle\rangle$. It is easy to see that the discriminant defines an isomorphism $I(F)/I^2(F) \xrightarrow{\sim} F^*/F^{*2}$. Since the 2-fold Pfister form $\langle\langle a, 1-a \rangle\rangle$ is trivial we have a well-defined ring homomorphism

$$K_*^M(F)/2 \rightarrow \prod_{n \geq 0} I^n(F)/I^{n+1}(F): \{a_1, \dots, a_n\} \mapsto \langle\langle a_1, \dots, a_n \rangle\rangle \bmod I^{n+1}(F),$$

which is surjective by the remarks above.

CONJECTURE 4.3 (MILNOR [20]). $K_n^M(F)/2 \xrightarrow{\sim} I^n(F)/I^{n+1}(F)$.

For $n \leq 2$ this conjecture was verified in [20].

Among the general properties of Milnor K -groups we will note the following:

(4.4) Let O be a discrete valuation ring with valuation v , fraction field F , and residue field k . Then there are canonical homomorphisms $\partial: K_n^M(F) \rightarrow K_{n-1}^M(k)$, which are completely characterized by the formula

$$\partial(\{x_1, \dots, x_n\}) = v(x_1) \cdot \{\bar{x}_2, \dots, \bar{x}_n\} \quad (x_2, \dots, x_n \in O^*)$$

[2].

REMARK 4.4.1. The following diagram commutes:

$$\begin{array}{ccc} K_n^M(F) & \rightarrow & K_n(F) \\ \partial \downarrow & & \downarrow \partial \\ K_{n-1}^M(k) & \rightarrow & K_{n-1}(k) \end{array}$$

² Remark added in proof: Ch. Wiebel pointed out to me that the map $K_n^M(Q) \rightarrow K_n(Q)$ is not injective for $n \geq 4$.

³ Remark added in proof: The formulation of this conjecture should be modified; in the present form it is easily seen to be false even in degree 4.

(4.5) *Transfer* (Bass and Tate [2], Kato [12]). If E/F is a finite extension, then it is possible to define canonical homomorphisms $N_{E/F}: K_n^M(E) \rightarrow K_n^M(F)$, which satisfy the usual properties of transfer and are completely characterized by the following reciprocity formula:

(4.5.1) If C/F is a complete regular curve, then for any $u \in K_{n+1}^M(C)$ one has $\sum_{x \in C} N_{F(x)/F}(\partial_x(u)) = 0$.

REMARK 4.5.2. The following diagram commutes:

$$\begin{array}{ccc} K_n^M(E) & \xrightarrow{N_{E/F}} & K_n^M(F) \\ \downarrow & & \downarrow \\ K_n(E) & \xrightarrow{N_{E/F}} & K_n(F) \end{array}$$

Milnor K -theory is closely related to homology properties of GL_n .

THEOREM 4.6 [29]. *Let F be an infinite field. Then the homomorphisms $H_n(\mathrm{GL}_n(F), Z) \rightarrow H_n(\mathrm{GL}_{n+1}(F), Z) \rightarrow \cdots \rightarrow H_n(\mathrm{GL}(F), Z)$ are isomorphisms. Moreover, the homology product*

$$F^* \otimes \cdots \otimes F^* = H_1(\mathrm{GL}_1(F)) \otimes \cdots \otimes H_1(\mathrm{GL}_1(F)) \rightarrow H_n(\mathrm{GL}_n(F))$$

defines an isomorphism

$$K_n^M(F) \rightarrow H_n(\mathrm{GL}_n(F))/H_n(\mathrm{GL}_{n-1}(F)) = H_n(\mathrm{GL}(F))/H_n(\mathrm{GL}_{n-1}(F)).$$

The last theorem provides a homomorphism

$$\begin{aligned} f: K_n(F) &= \pi_n(\mathrm{BGL}(F)^+) \rightarrow H_n(\mathrm{BGL}(F)^+) \\ &= H_n(\mathrm{GL}(F)) \rightarrow H_n(\mathrm{GL}(F))/H_n(\mathrm{GL}_{n-1}(F)) = K_n^M(F). \end{aligned}$$

PROPOSITION 4.5 [29]. (a) *The composition $K_n^M(F) \rightarrow K_n(F) \rightarrow K_n^M(F)$ coincides with multiplication by $(-1)^{n-1}(n-1)!$*

(b) *The composition $K_n(F) \rightarrow K_n^M(F) \rightarrow K_n(F)$ coincides with Chern class $c_{n,n}$.*

COROLLARY 4.7.1. *The kernel of the homomorphism $K_n^M \rightarrow K_n(F)$ is annihilated by $(n-1)!$.*

COROLLARY 4.7.2. *Suppose that O is a discrete valuation ring with fraction field F and residue field k . Then the diagram*

$$\begin{array}{ccc} K_3(F) & \xrightarrow{f} & K_3^M(F) \\ \partial \downarrow & & \downarrow \partial \\ K_2(F) & \xrightarrow{2} & K_2(F) \end{array}$$

commutes.

In connection with Milnor's conjecture (4.3) we will mention also Proposition 4.8.

PROPOSITION 4.8 [29]. *The image of the homomorphism $f: K_3(F) \rightarrow K_3^M(F)$ coincides with the kernel of Milnor's homomorphism*

$$K_3^M(F) \rightarrow I^3(F)/I^4(F).$$

5. K_3 and Bloch's group. For any field F denote by $D(F)$ the free abelian group with basis $[x]$ ($x \in F^* - 1$) and by $r: D(F) \rightarrow F^* \otimes F^*$ the homomorphism $[x] \mapsto x \otimes (1-x)$. There is an involution s on $F^* \otimes F^*$, given by $s(a \otimes b) = -(b \otimes a)$. It is easy to verify that the induced homomorphism $D(F) \rightarrow (F^* \otimes F^*)_s$ is trivial on elements of the form

$$[x] - [y] + \left[\frac{y}{x} \right] - \left[\frac{1-x^{-1}}{1-y^{-1}} \right] + \left[\frac{1-x}{1-y} \right] \quad (x \neq y \in F^* - 1).$$

We will denote by $T(F)$ the factor-group of $D(F)$ by the subgroup, generated by the above elements. The kernel of the induced homomorphism $T(F) \rightarrow (F^* \otimes F^*)_s$ is denoted $B(F)$ and is called the Bloch's group of F . Thus we have an exact sequence $0 \rightarrow B(F) \rightarrow T(F) \rightarrow (F^* \otimes F^*)_s \rightarrow K_2(F) \rightarrow 0$. For $x \neq 1$ put $\langle x \rangle = [x] + [x^{-1}]$; put also $\langle 1 \rangle = 0$.

LEMMA 5.1. (a) $x \mapsto \langle x \rangle$ defines a homomorphism $F^* \rightarrow {}_2T(F)$; in particular, $\langle x^2 \rangle = 0$.

(b) The element $c = [x] + [1-x] \in T(F)$ does not depend on the choice of $x \in F^* - 1$.

(c) $3c = \langle -1 \rangle$.

(d) If equation $x^2 + 1 = 0$ has solutions in F , then $3c = 0$; if equation $x^2 - x + 1 = 0$ has solutions in F , then $2c = 0$.

The group $B(F)$ has the following relation to $K_3(F)$. Denote by $\text{GM}(F)$ the subgroup of $\text{GL}(F)$, consisting of monomial matrices. This group is quasiperfect, so one can apply to $\text{BGM}(F)$ Quillen's plus-construction. The homotopy groups of $\text{BGM}(F)^+$ coincide in view of the Barrat-Priddy-Quillen theorem with stable homotopy groups of BF^* and hence are more or less understandable. The imbedding $\text{GM}(F) \hookrightarrow \text{GL}(F)$ induces a map $\text{BGM}(F)^+ \rightarrow \text{BGL}(F)^+$ and hence homomorphisms $\pi_i^s(BF^*) = \pi_i(\text{BGM}(F)^+) \rightarrow K_i(F)$. These homomorphisms are surjective in dimensions ≤ 2 .

THEOREM 5.2 [33]. *If the field F is infinite, then*

$$\text{coker}(\pi_3(\text{BGM}(F)^+) \rightarrow K_3(F)) = B(F)/2c.$$

The proof is done by means of homological methods. One proves first of all that

$$\text{coker}(\pi_3(\text{BGM}(F)^+) \rightarrow K_3(F)) = \text{coker}(H_3(\text{GM}(F)) \rightarrow H_3(\text{GL}(F))).$$

Next one computes $H_3(\text{GL}_2(F))/H_3(\text{GM}_2(F))$. This step is very close to the proof of Bloch's theorem [6]. Consider the complex $C_*(F)$ with $C_i(F)$ equal to the free abelian group, generated by $(i+1)$ -tuples (x_0, \dots, x_i) of distinct points of $\mathbf{P}^1(F)$. It is easy to see that all homology groups of this complex are zero,

except for H_0 , which is equal to Z . The natural action of $\mathrm{GL}_2(F)$ on $C_*(F)$ gives rise to a spectral sequence $H_p(\mathrm{GL}_2(F), C_q(F)) \Rightarrow H_{p+q}(\mathrm{GL}_2(F), Z)$. The action of $\mathrm{GL}_2(F)$ on the basis of $C_i(F)$ is transitive for $i = 0, 1, 2$ and the stabilizers of (0) , $(0, \infty)$, $(0, \infty, 1)$ are correspondingly equal to $B_2 = \begin{pmatrix} F^* & * \\ 0 & F^* \end{pmatrix}$, $T_2 = \begin{pmatrix} F^* & 0 \\ 0 & F^* \end{pmatrix}$, and F^* . Thus the term E^1 of the spectral sequence looks as follows:

$$\begin{array}{ccc} H_*(B_2) & H_*(T_2) & H_*(F^*) \\ & \coprod_{x \in \mathbb{P}^1(F) - \{0, \infty, 1\}} & \coprod_{x \neq y \in \mathbb{P}^1(F) - \{0, \infty, 1\}} \\ & & Z[x] \quad Z[x, y] \end{array}$$

where $[x]$ (resp. $[x, y]$) is the orbit of $(0, \infty, 1, x)$ (resp. $(0, \infty, 1, x, y)$). Using the fact that $H_*(B_2) = H_*(T_2)$, one computes easily the differential d_1 . The interesting E_2 -terms look as follows (s is the involution induced in $H_*(T_2)$ by the permutation of factors):

$$\begin{array}{cccc} H_3(T_2)_s & & & \\ H_2(T_2)_s = H_2(F^*) \oplus (F^* \otimes F^*)_s & (F^* \otimes F^*)_s & 0 & \\ F^* & 0 & 0 & \\ Z & 0 & 0 & T(F) \end{array}$$

The only nontrivial differential starting at $T(F)$ is the differential $d_3: T(F) \rightarrow H_2(F^*) \oplus (F^* \otimes F^*)_s$, which is given by the formula $d_3([x]) = x \wedge (1 - x) - x \otimes (1 - x) \in \Lambda^2(F^*) + (F^* \otimes F^*)_s$. It is clear that $E_{0,3}^\infty = \ker d_3 = B(F)$. Our spectral sequence defines a filtration on $H_3(\mathrm{GL}_2(F))$. We have also a filtration on $H_3(\mathrm{GM}_2(F))$ arising from the Hochschild-Serre spectral sequence, corresponding to the extension $1 \rightarrow T_2 \rightarrow \mathrm{GM}_2(F) \rightarrow S_2 \rightarrow 1$. In both cases the zero's term of filtration coincides with the image of $H_3(T_2)$. Thus the homomorphism $H_3(\mathrm{GM}_2(F)) \rightarrow H_3(\mathrm{GL}_2(F))$ takes $H_3(\mathrm{GM}_2(F))^0$ onto $H_3(\mathrm{GL}_2(F))^0$. Comparing the $E_{2,1}^2$ -terms of the spectral sequences under consideration, one shows easily that $H_3(\mathrm{GM}_2(F))^1$ is mapped onto $H_3(\mathrm{GL}_2(F))^1$. Finally $H_3(\mathrm{GM}_2(F)) = H_3(\mathrm{GM}_2(F))^1 + H_3(S_2)$ and the image of $H_3(S_2)$ in $H_3(\mathrm{GL}_2(F))$ is clearly contained in $H_3(T_2)$. Thus

$$H_3(\mathrm{GL}_2(F))/H_3(\mathrm{GM}_2(F)) = H_3(\mathrm{GL}_2(F))/H_3(\mathrm{GL}_2(F))^1 = E_{0,3}^\infty = B(F).$$

Next one checks that the kernel of $H_3(\mathrm{GL}_2(F)) \rightarrow H_3(\mathrm{GL}_3(F))$ is contained in the image of $H_3(\mathrm{GM}_2)$. After that one has only to compute the intersection of $H_3(\mathrm{GL}_2(F))$ and $H_3(\mathrm{GM}(F))$ in $H_3(\mathrm{GL}(F))$. In view of the isomorphism $H_3(\mathrm{GL}(F))/H_3(\mathrm{GL}_2(F)) = K_3^M(F)$, this is equivalent to the computation of the kernel of $H_3(\mathrm{GM}(F)) \rightarrow K_3^M(F)$. The answer is as follows: consider on $H_3(\mathrm{GM}(F))$ the filtration, arising from the Hochschild-Serre spectral sequence; then $H_3(\mathrm{GM}(F))^2 \cap H_3(\mathrm{GL}_2(F)) = H_3(\mathrm{GM}_2(F))$. Since $H_3(\mathrm{GM}(F)) = H_3(\mathrm{GM}(F))^2 + H_3(S)$ we deduce that

$$H_3(\mathrm{GL}(F))/H_3(\mathrm{GM}(F)) = B(F)/\mathrm{Im}(H_3(S))$$

and it is sufficient to check now that $\mathrm{Im}(H_3(S)) = 2c$.

To apply Theorem 5.2 it is necessary to know the group $\pi_3(\mathrm{BGM}(F)^+)$ and its image in $K_3(F)$. Using the spectral sequence $H_i(F^*, \pi_j^s(\mathrm{pt})) \Rightarrow \pi_{i+j}^s(BF^*)$ one easily proves Proposition 5.3.

PROPOSITION 5.3. Denote by $\overline{K}_3^M(F)$ the image of $K_3^M(F)$ in $K_3(F)$ and by G (resp. $G\mu$) the subgroup of GM , consisting of monomial matrices with entries ± 1 (resp. with entries from the group μ of roots of unity).

(a) $\text{Im}(\pi_3(\text{BGM}(F)^+) \rightarrow K_3(F)) = \overline{K}_3^M(F) + \text{Im}((\pi_3(BG\mu)^+) \rightarrow K_3(F))$.

(b) There is a canonical surjective homomorphism $\text{Tor}(\mu, \mu) = \text{Tor}(F^*, F^*) \rightarrow \text{Im}(\pi_3(\text{BGM}(F)^+)/K_3^M(F) + \text{Im}(\pi_3(BG^+)))$.⁴

COROLLARY 5.3.1. The group $B(F)$ does not change under purely transcendental extensions.

COROLLARY 5.3.2. Denote by F_0 the subfield of constants in F . Then $K_3(F)/K_3(F_0) + K_3^M(F) = B(F)/B(F_0)$.

In the case of K_3 , Conjecture 4.1 may be specified as follows:

CONJECTURE 5.4. $B(F) = B(F_0)$.

We will need below two slightly different descriptions of $B(F)$.

(5.5) Denote by $\overline{F^*} \otimes \overline{F^*}$ the factor group of $F^* \otimes F^*$ by the subgroup generated by elements $a \otimes (-a)$, and by $T'(F)$ the factorgroup of $T(F)$ by the subgroup generated by elements $\langle a \rangle$. Since $r(\langle a \rangle) = a \otimes (-a)$ we get an induced homomorphism $T'(F) \rightarrow \overline{F^*} \otimes \overline{F^*}$, whose kernel we will denote by $B'(F)$.

LEMMA 5.5.1. $B'(F) = B(F)/\langle -1 \rangle$.

(5.6) One can describe $B(F)$ equally in terms of relations on $a \otimes (1-a)$ directly in $F^* \otimes F^*$. It is easy to verify that

$$\begin{aligned} r \left([x] - [y] + \left[\frac{y}{x} \right] - \left[\frac{1-x^{-1}}{1-y^{-1}} \right] + \left[\frac{1-x}{1-y} \right] \right) \\ = x \otimes \frac{1-x}{1-y} + \frac{1-x}{1-y} \otimes x = r \left(\left\langle x \frac{1-x}{1-y} \right\rangle - \langle x \rangle - \left\langle \frac{1-x}{1-y} \right\rangle \right). \end{aligned}$$

Thus the kernel of r contains elements

$$\begin{aligned} [x] - [y] + \left[\frac{y}{x} \right] - \left[\frac{1-x^{-1}}{1-y^{-1}} \right] + \left[\frac{1-x}{1-y} \right] \\ - \left\langle x \frac{1-x}{1-y} \right\rangle + \langle x \rangle + \left\langle \frac{1-x}{1-y} \right\rangle_{(x \neq y \neq F^* - 1)}, \\ \langle xyz \rangle - \langle xy \rangle - \langle xz \rangle - \langle yz \rangle + \langle x \rangle + \langle y \rangle + \langle z \rangle, \\ \langle x^2 \rangle - 4\langle x \rangle \end{aligned}$$

(where, as always, $\langle 1 \rangle = 0$). Denote by $T''(F)$ the factor group of $D(F)$ by the above elements, and by $B''(F)$ the kernel of the homomorphism $T''(F) \rightarrow F^* \otimes F^*$.

⁴Remark added in proof: More precisely, the relation between $K_3(F)$ and $B(F)$ is given by the exact sequence $0 \rightarrow \text{Tor}(F^*, F^*) \sim \rightarrow K_3(F)_{\text{ind}} \rightarrow B(F) \rightarrow 0$ where $K_3(F)_{\text{ind}} = K_3(F)/K_3^M(F)$ and $\text{Tor}(F^*, F^*) \sim$ is the unique nontrivial extension of $\mathbb{Z}/2$ by means of $\text{Tor}(F^*, F^*)$.

LEMMA 5.6.1. $B''(F) = B(F)$.

6. Divisibility in Bloch's group.⁵ All fields considered in this section are supposed to contain an algebraically closed subfield.

Suppose that F is a discretely valued field with valuation ring O and residue field k . Choose a local parameter π . This choice defines a homomorphism $s_\pi: F^* \rightarrow k^*: x \mapsto x/\pi^{v(x)}$ and induced homomorphisms $F^* \otimes F^* \rightarrow k^* \otimes k^*$, $\overline{F^* \otimes F^*} \rightarrow \overline{k^* \otimes k^*}$. We will take $\bar{x} = \infty$ for $x \notin O$ and we will define elements $[0], [\infty], [1] \in T'(k)$ as zero.

LEMMA 6.2. $[x] \mapsto [\bar{x}]$ defines a homomorphism $T'(F) \xrightarrow{s} T'(k)$. Moreover, the following diagram commutes:

$$\begin{array}{ccc} T'(F) & \rightarrow & \overline{F^* \otimes F^*} \\ s \downarrow & & s_\pi \downarrow \\ T'(k) & \rightarrow & \overline{k^* \otimes k^*} \end{array}$$

Hence s takes $B'(F)$ to $B'(k)$.

If E/F is a finite extension, then $N_{E/F}: K_3(E) \rightarrow K_3(F)$ defines in view of (5.3) and (4.5) the transfer $N_{E/F}: B(E) \rightarrow B(F)$. A slight modification of the proof of (2.3) now gives

PROPOSITION 6.2. Let C be a smooth connected curve over an algebraically closed field F . For any two points $x, y \in C$ the specialization homomorphisms $s_x, s_y: B(F(C)) \rightarrow B(F)$ coincide on B/n and ${}_n B$.

THEOREM 6.3. If F is algebraically closed, then $B(F)$ is uniquely divisible.

Since $\overline{F^* \otimes F^*}$ and $K_2(F)$ are uniquely divisible, it is sufficient to prove the unique divisibility of $T'(F) = T(F)$. The divisibility of $T(F)$ was proved in [6]. It follows from the formulae $[x^p] = p(\sum_{\xi \in \mu_p} [\xi x])$ ($p \neq \text{char } F$), $[x^p] = p^2[x]$ ($p = \text{char } F$), which are valid for any field F . To prove these formulae consider the element $[t^p] - p(\sum_{\xi \in \mu_p} [\xi t]) \in B(F(t))$. Since $B(F(t)) = B(F)$ this element coincides with any of its specializations. But specializing at zero we get zero. Now, to prove the unique divisibility, we will define a homomorphism $T'(F) \rightarrow T'(F)$ inverse to multiplication by p by means of the formula $[x] \rightarrow \sum_{y^p=x} [y]$. We have to check that the defining relation on $[x]$ goes to zero, i.e., to check the formula (where $u, v, w \notin \mu_p \cup 0$ and $w^p = (1 - u^p)/(1 - v^p)$):

$$\begin{aligned} & \sum_{\xi \in \mu_p} [\xi u] - \sum_{\xi \in \mu_p} [\xi v] + \sum_{\xi \in \mu_p} [\xi \cdot v/u] \\ & - \sum_{\xi \in \mu_p} [\xi \cdot wv/u] + \sum_{\xi \in \mu_p} [\xi \cdot w] = 0. \end{aligned}$$

Note that this element lies in ${}_p B(F)$. Now fix w and consider the curve C , given by equation $(1 - V^p)w^p = 1 - U^p$. We can consider the universal element in

⁵ Remark added in proof: A different and much more powerful approach to the study of divisibility in $K_3(F)_{\text{ind}}$ (and hence in $B(F)$) is developed in [46].

${}_p B(F(C))$ for which the element under consideration is a specialization. Specializing this element in the point $U = 1, V = 1$ we will get zero and it is sufficient to apply Proposition 6.2.

Suppose from now on that $\text{char } F \neq 2$. Let E/F be a quadratic extension: $E = F(\alpha)$, $\alpha^2 = a \in F^*$. Denote by A the image of $E^* \otimes F^* \rightarrow E^* \otimes E^*$. One sees easily that $N_{E/F} \otimes \text{id}: E^* \otimes F^* \rightarrow F^* \otimes F^*$ induce a homomorphism $N: A \rightarrow F^* \otimes F^*$. Let $T''(E/F)$ denote the inverse image of A in $T''(E)$. It is not difficult to find generators and relations for this group. The important point is that relations are of "rational character" (i.e., may be parametrized by means of F -rational varieties).

PROPOSITION 6.4. *There exists a canonical homomorphism (given by rational formulae) $L_{E/F}: T''(E/F) \rightarrow T''(F)$, making the following diagram commutative*

$$\begin{array}{ccc} T''(E/F) & \rightarrow & A \\ L_{E/F} \downarrow & & N \downarrow \\ T''(F) & \rightarrow & F^* \otimes F^* \end{array}$$

and hence inducing the homomorphism $L_{E/F}: B(E) \rightarrow B(F)$.

$L_{E/F}$ is given explicitly on generators. To check that relations go to zero one remarks that the image of any relation is a rationally parametrized element of $B(F)$ and hence should be zero.

THEOREM 6.5. *Denote by t the generator of $\text{Gal}(E/F)$. The following sequence is exact: $B(E) \xrightarrow{1-t} B(E) \xrightarrow{L_{E/F}} B(F)$.*

As in the proof of (1.5) we reduce, first of all, the general case to the case where $N_{E/F}: E^* \rightarrow F^*$ is surjective. In the present situation this is trivial: to make $b \in F^*$ a norm it is sufficient to pass to the function field on a conic C with equation $X^2 - aY^2 = b$. The field $E(C)$ is purely transcendental over E and hence $B(E(C)) = B(E)$. Thus

$$\ker L_{E/F}/(1-t)B(E) \hookrightarrow \ker L_{E(C)/F(C)}/(1-t)B(E(C)).$$

Supposing now that $N_{E/F}: E^* \rightarrow F^*$ is surjective we define a map $f: T''(F) \rightarrow T''(E/F)_t$ by means of the formula

$$f([x]) = [-z] + [(1+z^t)z/(1+z)] - [-(1+z^t)/(1+z)],$$

where $z \in E^*$ is such that $N_{E/F}(z) = x$ and $\text{Tr}(z) \neq -2$. One verifies then that f and $L_{E/F}$ are mutually inverse. Set $M = \text{Im}(T''(E/F) \rightarrow A)$. We get two short exact sequences of t -modules:

$$0 \rightarrow B(E) \rightarrow T''(E/F) \rightarrow M \rightarrow 0, \quad 0 \rightarrow M \rightarrow A \rightarrow K_2(E) \rightarrow 0.$$

One computes easily the homology groups of $G = \text{Gal}(E/F)$ with coefficients in A and $K_2(F)$ (in the second case using essentially the results of §1). This makes it possible to compute homology with coefficients in M : $H_i(G, M) = \mathbb{Z}/2$. Now

it is easy to check that $H_1(G, T''(E/F)) \rightarrow H_1(G, M)$ is surjective and hence $(1-t)T''(E/F) \cap B(E) = (1-t)B(E)$.

Applying (6.5) to the universal Kummer Extension we get

COROLLARY 6.6. *For any F (containing an algebraically closed subfield) of characteristics $\neq 2$ the group $B(F)$ does not have 2-torsion.*

THEOREM 6.7. *In conditions of (6.6) the group $B(F)$ is uniquely 2-divisible.*⁶

(6.7.1) Let C be a conic over F . If $B(F)$ is 2-divisible then $B(F(C))$ is also 2-divisible.

Let E be a quadratic extension of F splitting C . If $u \in B(F(C))$ then $2u = N_{E(C)/F(C)}(u_{E(C)}) = N_{E/F}(u_{E(C)})_{F(C)}$ (since $B(E(C)) = B(E)$). We can write $N_{E/F}(u_{E(C)}) = 4v$ and hence $u = 2v_{F(C)}$.

COROLLARY 6.7.2. *If F is a function field on a product of conics defined over an algebraically closed field, then $B(F)$ is 2-divisible.*

Using the description of the 2-torsion in K_2 one can easily prove the exactness of the sequence $0 \rightarrow B(F)/2 \rightarrow T'(F)/2 \rightarrow \Lambda^2(F^*/F^{*2})$. This makes it possible to write down the universal elements of $B/2$. Fields of definition of these universal elements are function fields on products of conics, so we deduce from (6.7.2) that these universal elements are zero. The specialization argument finishes the proof.

COROLLARY 6.8. *Let F be as above and let F_0 denote its subfield of constants. Then $K_3(F) = K_3(F_0) + K_3^M(F) + 2K_3(F)$.*

COROLLARY 6.9. *For F as above, $K_3^M(F)/2 \xrightarrow{\sim} I^3(F)/I^4(F)$.*

In fact, $\ker(K_3^M(F) \rightarrow I^3/I^4) = \text{Im}(K_3(F) \rightarrow K_3^M(F))$ but the image of all three terms, which appear in (6.8), is clearly contained in $2K_3^M(F)$.

COROLLARY 6.10. *Let F be as above and let X/F be a smooth variety. Then the images of $(K_3(F(X)) \rightarrow \coprod_{\text{codim } x=1} K_2(F(x)))$ and $(K_3^M(F(X)) \rightarrow \coprod_{\text{codim } x=1} K_2(F(x)))$ coincide.*⁷

REMARK 6.11. It seems that (6.10) together with Merkurjev's Theorem 3.3 are sufficient to prove Hilbert's Theorem 90 for K_3^M (for quadratic extensions) and, in particular, to prove that $K_3^M(F)/2 = H^3(F, \mu_2)$; but we have not yet checked all the details.⁸

7. Higher Chow groups. To visualize the relations between K -theory and étale cohomology Beilinson [3] conjectured the existence of a certain "universal" cohomology theory on the category of schemes, which is directly related both to K -theory and to étale cohomology (this theory should be analogous to the

⁶ *Remark added in proof:* The group $B(F)$ is uniquely divisible for any F , containing an algebraically closed subfield [46].

⁷ *Remark added in proof:* Statements 6.9 and 6.10 are true for any field F [46].

⁸ *Remark added in proof:* Hilbert's Theorem 90 for $K_3^M(F)/2 = H^3(F, \mu_2)$ are proved in [45, 48].

integral singular cohomology theory in topology). A closely related list of conjectures was proposed by Lichtenbaum [15]. According to Beilinson there should exist complexes of sheaves $\Gamma(i)$ on the big Zariski site, satisfying (among others) the following properties:

- (a) $\Gamma(i) = 0$ for $i < 0$, $\Gamma(0) = Z$, $\Gamma(1) = O^*[-1]$.
- (b) For $i \geq 1$ the complex $\Gamma(i)$ is acyclic outside $1, \dots, i$; for a smooth X , $H^i(\Gamma(i))$ coincides with the sheaf of Milnor groups K_i^M .
- (c) For any n invertible on a "good" smooth X one has $\Gamma(i) \otimes^L Z/n = \tau_{\leq i} R\pi_* Z/n[i]$, where $\pi: X_{\text{et}} \rightarrow X_{\text{zar}}$ is the canonical morphism.
- (d) There exists a spectral sequence $H^i(X, \Gamma(j)) \Rightarrow K'_{2j-i}(X)$, which is split up to standard factorials by means of Chern classes. The resulting filtration on K' -theory coincides with γ -filtration.

Recently Bloch has constructed a theory which satisfies properties (a) and (d). There is no doubt that this is the expected theory, but it is very difficult to attack the remaining properties. We will work in the category of quasiprojective varieties over a field. Define the standard simplex Δ^n as a hyperplane in A^{n+1} , defined by the equation $t_0 + \dots + t_n = 1$. Δ^n form a cosimplicial variety. For any variety X define $z^i(X, n)$ to be a subgroup in the group $Z^i(X \times \Delta^n)$, consisting of those cycles which properly intersect $X \times \Delta^m$ for any face $\Delta^m \subset \Delta^n$. $z^i(-, -)$, is clearly a complex (of degree -1) of sheaves in any reasonable topology; the expected complex $\Gamma(i)$ is obtained from $z^i(-, -)$ by reindexing. Bloch set $\text{CH}^i(X, n)$ equal to the n th homology group of $z^i(X, -)$ (the group $\text{CH}^i(X, 0)$ coincides evidently with Chow groups of cycles of codimension i modulo rational equivalence). The groups $\text{CH}^i(X, n)$ are contravariant functors of X with respect to flat maps; one can define the inverse image on $\text{CH}^i(X, n)$ with respect to arbitrary maps if one restricts to the subcategory of smooth varieties. The groups $\text{CH}^i(X, n)$ are covariant functors of X with respect to proper maps. Bloch has also proved the following properties of $\text{CH}^i(X, n)$:

(7.1) $\text{CH}^i(X, n)$ are homotopy invariant: $\text{CH}^i(X, n) = \text{CH}^i(X \times A^1, n)$.

(7.2) *Localization.* If $Y \subset X$ is a closed subvariety of pure codimension d , then there is an exact sequence $\text{CH}^i(X - Y, n + 1) \rightarrow \text{CH}^{i-d}(Y, n) \rightarrow \text{CH}^i(X, n) \rightarrow \text{CH}^i(X - Y, n) \rightarrow \dots \rightarrow \text{CH}^i(X - Y, 0) \rightarrow 0$.

(7.3) *Products.* For any X, Y there are canonical pairings $\text{CH}^i(X, n) \otimes \text{CH}^j(Y, n) \rightarrow \text{CH}^{i+j}(X \times Y, n + m)$. Combining these products with inverse image along the diagonal one gets on $\text{CH}^*(X, *)$ (for X smooth) a structure of bigraded ring.

(7.4)

$$\text{CH}^1(X, q) = \begin{cases} \text{Pic } X, & q = 0, \\ \Gamma(X, \sigma_X^*), & q = 1, \\ 0, & q = 2. \end{cases}$$

(7.4) *Relations to K-theory.* Bloch shows that $\text{CH}^*(X, *)$ satisfy the Gillet axioms [44] and hence there is a theory of Chern classes with values in $\text{CH}^*(X, *)$.

He proves that these Chern classes define isomorphisms

$$\mathrm{CH}^i(X, n) \otimes Q = \mathrm{gr}^i K'_n(X) \otimes Q.$$

(7.5) Gersten conjecture is true for $\mathrm{CH}^i(X, n)$.

The spectral sequence relating higher Chow groups to K -theory was constructed earlier by Landsburg [13].

Consider the case of a field. It is clear that $\mathrm{CH}^i(\mathrm{Spec} F, n) = 0$ when $n < i$. One can easily verify also that $\mathrm{CH}^n(\mathrm{Spec} F, n) = K_n^M(F)$. These facts correspond exactly to the properties conjectured by Beilinson in (b) above. However, the remaining point of (b) means that $\mathrm{CH}^i(\mathrm{Spec} F, n) = 0$ when $n \geq 2i$. This seems to be an extremely difficult question for $i > 1$. Finally, we mention that the group $\mathrm{CH}^2(\mathrm{Spec} F, 3)$ coincides with $K_3(F)/K_3^M(F)$ and thus is very close to Bloch's group $B(F)$.

8. Etale K -theory. For any simplicial scheme X one can construct a certain pro-space X_{et} —its etale topological type [1, 8]. $X \mapsto X_{\mathrm{et}}$ is a functor from schemes to pro-spaces. The main property of X_{et} is that its fundamental group coincides with the fundamental group of X as defined by Grothendieck, and its cohomology groups with finite coefficients coincide with etale cohomology groups of X .

For a variety over C its etale K -theory may be defined as complex K -theory of the pro-space X_{et} . In the general case, one can proceed as follows [7]. Fix a prime integer l and denote $Z[l^{-1}]$ by R . We will consider only schemes over R . For any X one has morphisms of pro-spaces

$$X_{\mathrm{et}} \rightarrow (\mathrm{Spec} R)_{\mathrm{et}} \leftarrow (\mathrm{BGL}_n)_{\mathrm{et}}.$$

Consider now the space of relative l -adic functions [7, 8]

$$\mathrm{Hom}_l(X_{\mathrm{et}}, (\mathrm{BGL}_n)_{\mathrm{et}})_{R_{\mathrm{et}}}$$

and set

$$K_i^{\mathrm{et}}(X) = \varinjlim_n \pi_i(\mathrm{Hom}_l(X_{\mathrm{et}}, (\mathrm{BGL}_n)_{\mathrm{et}})_{R_{\mathrm{et}}})$$

and

$$K_i^{\mathrm{et}}(X, Z/l^\nu) = \varinjlim_n \pi_i(\mathrm{Hom}(X_{\mathrm{et}}, (\mathrm{BGL}_n)_{\mathrm{et}})_{R_{\mathrm{et}}}, Z/l^\nu).$$

Etale K -theory is easy to compute in view of spectral sequences relating it to etale cohomology (which are strongly convergent if X has finite l -cohomological dimension)

$$\begin{aligned} E_2^{p,q} &= H_{\mathrm{cont}}^p(X_{\mathrm{et}}, Z_l(q/2)) \Rightarrow K_{q-p}^{\mathrm{et}}(X), \\ E_2^{p,q} &= H^p(X_{\mathrm{et}}, Z/l^\nu(q/2)) \Rightarrow K_{q-p}^{\mathrm{et}}(X, Z/l^\nu), \\ & (E_2^{p,q} \text{ is zero if } q \text{ is odd}). \end{aligned}$$

For X quasiprojective over a noetherian R -algebra there are natural maps $K_i(X) \rightarrow K_i^{\mathrm{et}}(X)$, $K_i(X, Z/l^\nu) \rightarrow K_i^{\mathrm{et}}(X, Z/l^\nu)$.

THEOREM 8.1 [7]. *Let A be the ring of integers in an algebraic number field F and let l be a prime integer. If $l = 2$ assume, in addition, that $F \ni \sqrt{-1}$. Then the natural map*

$$K_i(A) \otimes \mathbb{Z}_l \rightarrow K_i^{\text{et}}(A[1/l])$$

is surjective.

It should be mentioned that Quillen-Lichtenbaum conjectures for number fields are equivalent to the fact that the homomorphism considered in (8.1) is bijective.

Let A be an R -algebra, containing a primitive l^ν th root of unity ξ . The group $\pi_2(BA^*, \mathbb{Z}/l^\nu)$ coincides with the group ${}_l A$ of l^ν th roots of unity in A . The image of ξ under a canonical homomorphism $\pi_2(BA^*, \mathbb{Z}/l^\nu) \rightarrow K_2(A, \mathbb{Z}/l^\nu)$ (induced by the evident morphism $BA^* \rightarrow \text{BGL}(A)^+$) is denoted by β and is called the Bott element. Let X be a scheme of finite l -cohomological dimension over A . It is easy to see that étale K -theory of X , $K^{\text{et}}(X, \mathbb{Z}/l^\nu)$, is 2-periodical and this periodicity is given by multiplication by β . Thus we get an induced map $K_*(X, \mathbb{Z}/l^\nu)[\beta^{-1}] \rightarrow K_*^{\text{et}}(X, \mathbb{Z}/l^\nu)$ (one has to be more careful when $l = 2$ or 3 since in these cases there are difficulties with ring structure on $K_*(X, \mathbb{Z}/l^\nu)$). The fundamental result, relating algebraic and étale K -theory is the following theorem of Thomason [38].

THEOREM 8.2. *Under mild additional hypotheses (see [38] for the exact formulation) the induced map*

$$K_*(X, \mathbb{Z}/l^\nu)[\beta^{-1}] \rightarrow K_*^{\text{et}}(X, \mathbb{Z}/l^\nu)$$

is an isomorphism.

REFERENCES

1. M. Artin and B. Mazur, *Étale homotopy*, Lecture Notes in Math., Vol. 100, Springer-Verlag, Berlin-New York, 1969.
2. H. Bass and J. Tate, *The Milnor ring of a global field*, Lecture Notes in Math., Vol. 342, Springer-Verlag, Berlin-New York, 1973, pp. 349–446.
3. A. Beilinson, Letter to C. Soulé, November 1, 1982.
4. S. Bloch, *Torsion algebraic cycles, K_2 and Brauer groups of function fields*, Bull. Amer. Math. Soc. **80** (1974), 941–945.
5. ———, *Algebraic cycles and higher K -theory*, Preprint, 1985.
6. J. Dupont and C. Sah, *Scissor congruences. II*, J. Pure Appl. Algebra **25** (1982), 159–195.
7. W. Dwyer and E. Friedlander, *Algebraic and étale K -theory*, Preprint, 1983.
8. E. Friedlander, *Étale homotopy of simplicial schemes*, Ann. of Math. Studies, No. 104, Princeton Univ. Press, Princeton, N.J., 1982.
9. S. Gersten, *Problems about higher K -functors*, Lecture Notes in Math., Vol. 341, Springer-Verlag, Berlin-New York, 1973, pp. 43–57.
10. H. Gillet and R. Thomason, *The K -theory of strict hensel local rings and a theorem of Suslin*, J. Pure Appl. Algebra **34** (1984), 241–254.
11. A. Juffrjakov, *K_2 of a local division algebra*, Dokl. Akad. Nauk. SSSR (to appear).
12. K. Kato, *A generalisation of local class field theory by using K -groups*, J. Fac. Sci. Univ. Tokyo 1A Math. **26** (1979), 303–376; **27** (1980), 603–683.
13. S. Landsburg, *Relative cycles and algebraic K -theory*, Preprint, 1983.

14. S. Lichtenbaum, *Values of zeta-function, etale cohomology and algebraic K-theory*, Lecture Notes in Math., Vol. 342, Springer-Verlag, Berlin-New York, 1973, pp. 489–501.
15. —, *Values of zeta-function at non-negative integers*, Lecture Notes in Math., Vol. 1068, Springer-Verlag, Berlin-New York, 1984, pp. 127–138.
16. A. Merkurjev, *On the norm-residue symbol of degree two*, Dokl. Akad. Nauk SSSR **261** (1981), 542–547; English transl. in Soviet Math. Dokl. **24** (1981).
17. A. Merkurjev and A. Suslin, *K-cohomology of Severi-Brauer varieties and norm-residue homomorphism*, Izv. Akad. Nauk SSSR **46** (1982), 1011–1046; English transl. in Math. USSR-Izv. **21** (1983), 307–340.
18. A. Merkurjev, *The group SK_2 for quaternion algebras*, Dokl. Akad. Nauk SSSR (to appear).
19. J. Milnor, *Algebraic K-theory*, Ann. of Math. Studies, No. 72, Princeton Univ. Press, Princeton, N.J., 1971.
20. —, *Algebraic K-theory and quadratic forms*, Invent. Math. **9** (1970), 318–344.
21. —, *On the homology of Lie groups made discrete*, Comment. Math. Helv. **58** (1983), 72–85.
22. D. Quillen, *On the cohomology and K-theory of the general linear group over finite fields*, Ann. of Math. **96** (1972), 552–586.
23. —, *Higher algebraic K-theory. I*, Lecture Notes in Math., Vol. 341, Springer-Verlag, Berlin-New York, 1973, pp. 85–147.
24. —, *Higher algebraic K-theory*, Proc. Internat. Congr. Math. (Vancouver, B.C., 1974), Vol. 1, Canad. Math. Congress, Montreal, Que., 1975, pp. 171–177.
25. V. Schekhtman, *The Riemann-Roch theorem and the Atiyah-Hirzebruch spectral sequence*, Uspekhi Mat. Nauk **35** (1980), 179–180; English transl. in Russian Math. Surveys **35** (1980).
26. J.-P. Serre, *Groups algébriques et corps de classes*, Hermann, Paris, 1959.
27. A. Suslin, *The quaternion homomorphism for the function field on a conic*, Dokl. Akad. Nauk SSSR **265** (1982), 292–296; English transl. in Soviet Math. Dokl. **26** (1982), 72–77.
28. —, *Torsion in K_2 of fields*, LOMI Preprint, E282, 1982.
29. —, *Homology of GL_n , characteristic classes and Milnor K-theory*, Lecture Notes in Math., Vol. 1046, Springer-Verlag, Berlin-New York, 1984, pp. 357–375.
30. —, *Algebraic K-theory and norm-residue homomorphism*, Itogi Nauki i Tekhniki, Sovr. Probl. Matem. **25** (1984), 115–209.
31. —, *On the K-theory of algebraically closed fields*, Invent. Math. **73** (1983), 241–245.
32. —, *On the K-theory of local fields*, J. Pure Appl. Algebra **34** (1984), 301–318.
33. —, *K_3 of a field and Bloch's group* (to appear).
34. —, *Divisibility in Bloch's group and Milnor's conjecture on quadratic forms* (not to appear).
35. A. Suslin and A. Jaffrjakov, *On the K-theory of local division algebras*, Dokl. Akad. Nauk SSSR (to appear).
36. R. Swan, *K-theory of quadratic hypersurfaces*, Ann. of Math. **122** (1985).
37. J. Tate, *On the torsion in K_2 of fields*, Proc. Internat. Sympos. Algebr. Number Theory (Kyoto, 1976), Japan Soc. for the Prom. of Science, Tokyo, 1977.
38. R. Thomason, *Algebraic K-theory and etale cohomology*, Ann. Sci. École Norm. Sup. (4) **18** (1985), 437–552.
39. J.-L. Colliot-Thélène, *Hilbert's theorem 90 for K_2 , with application to the Chow groups of rational surfaces*, Invent. Math. **71** (1983), 1–20.
40. J.-L. Colliot-Thélène, J.-J. Sansuc, and C. Soulé, *Torsion dans le groupe de Chow de codimension deux*, Duke Math. J. **50** (1983), 763–801.
41. J. Murre, *Applications of algebraic K-theory to the theory of algebraic cycles*, Lecture Notes in Math., Vol. 1124, Springer-Verlag, Berlin-New York, 1985.
42. J. Jardine, *Simplicial objects in a Grothendieck topos*, Preprint, 1983.
43. I. Panin, *On the Hurewicz theorem and the K-theory of complete discrete valuation rings*, Izv. Akad. Nauk SSSR **50** (1986).

- 44. H. Gillet, *Riemann-Roch theorems for higher algebraic K-theory*, Adv. in Math. **40** (1981), 203–289.
 - 45. A. Merkurjev and A. Suslin, *On the norm-residue homomorphism of degree three*, LOMI Preprint E-9-86, 1986.
 - 46. —, *On the K_3 of a field*, LOMI Preprint, E-2-87, 1987.
 - 47. M. Rost, *Injectivity of $K_2D \rightarrow K_2F$ for quaternion algebras*, Preprint, Regensburg, May 1986.
 - 48. —, *Hilbert 90 for K_3 for degree-two extensions*, Preprint, Regensburg, May 1986.
- LOMI, LENINGRAD 191011, USSR

Representations of Reductive Lie Groups

DAVID A. VOGAN, JR.

1. Introduction. Representation theory is built around two of the fundamental ideas of mathematics: symmetry and linearization. Unfortunately, only one of these is an idea which we would be proud to bring home and introduce to our parents. The development of the subject has often paralleled these dual origins, with insights of breathtaking power and beauty hidden within a maze of technical preliminaries. In this paper, I hope to sketch the development of one of these insights (the Kirillov-Kostant philosophy of coadjoint orbits) as it applies to representations of reductive Lie groups. The technical preliminaries, along with their own more difficult beauties, I will largely neglect.

To begin, suppose G is a group. To regard G as a group of symmetries means to consider a set X on which G acts. That is, for each element g of G there is a permutation $x \rightarrow g \cdot x$ of X ; and the permutation associated to the product of two group elements is the product of the permutations associated to the elements. Formally,

DEFINITION 1.1. An *action* or *permutation representation* of a group G is a pair (σ, X) , with X a set, and σ a homomorphism from G to the group of permutations of X . When no confusion results, we write $g \cdot x$ instead of $\sigma(g)(x)$. We sometimes say that X is a G -space.

Here is another formulation of the same idea.

DEFINITION (1.1)'. An action of a group G on a set X is a map $G \times X \rightarrow X$, $(g, x) \rightarrow g \cdot x$, satisfying the following conditions:

- (a) for all g and h in G , and x in X , $g \cdot (h \cdot x) = (gh) \cdot x$; and
- (b) if e is the identity element of G , then $e \cdot x = x$.

The most important example is this: if H is any subgroup of G , then G acts on the set G/H of left cosets of H in G , by $g \cdot (xH) = (gx)H$.

There are a few interesting things to say about actions even in this generality.

DEFINITION 1.2. Suppose the group G acts on the set X . Fix an element x of X . The *orbit* of x under the action is the subset

$$G \cdot x = \{g \cdot x | g \in G\} \subset X.$$

The author is an Alfred P. Sloan Foundation Fellow. This work is supported, in part, by National Science Foundation Grant DMS-8504029.

The *isotropy group* of the action at x is the subgroup

$$G(x) = \{g \in G \mid g \cdot x = x\} \subset G.$$

The action is said to be *transitive* if X consists of exactly one orbit. In that case, we say that X is a *homogeneous space* for G .

Any group acts in a unique way on the empty set. Since the number of orbits is zero, the action is not transitive.

LEMMA 1.3. *Suppose G acts on X . Then X is the disjoint union of all the orbits of X on G . This is the unique decomposition of X into a union of homogeneous spaces for G .*

This lemma shows that transitive actions have a particular importance. To describe all of them, we need to know what an equivalence between two actions is.

DEFINITION 1.4. Suppose G acts on two sets X and Y . A map f from X to Y is called *G -equivariant* if it respects the action of G : $f(g \cdot x) = g \cdot f(x)$. The actions are called *equivalent* if there is a G -equivariant bijection from X to Y .

LEMMA 1.5. *Suppose G acts on X , and $x \in X$. Then there is a well-defined map from the coset space $G/G(x)$ to the orbit $G \cdot x$ (Definition 1.2), given by $gG(x) \rightarrow g \cdot x$. This map is an equivariant bijection from $G/G(x)$ onto the orbit $G \cdot x$.*

Suppose G acts transitively on X and Y ; fix points x and y in these sets. Then X is equivalent to Y if and only if the isotropy groups $G(x)$ and $G(y)$ are conjugate as subgroups of G .

In one sense, these results describe group actions quite completely. It is misleading to take them too seriously, however. To know no more of a sphere than that it is equivalent to $\mathrm{SO}(3)/\mathrm{SO}(2)$ is not to know enough.

Almost any kind of additional structure can be imposed on Definition (1.1)'. For example, suppose G is a Lie group, and X is a manifold. Then an action of G of X is called *smooth* if the map from $G \times X$ to X is a smooth map. The questions one wants to ask about a group action usually depend on some such additional structure. For example, if G and X are finite, one can ask for the cardinality of X . If they are topological, one can examine the topological type of X . If they are algebraic varieties, then the singularities of X may be of interest.

All of these questions—and the theory of group actions generally—are nonlinear, in the simple-minded sense that they do not involve vector spaces. Therefore they are quite hard. Some hint of this can be seen even in Lemma 1.5: finding all the subgroups of a group is an enormously difficult task, even for such apparently innocuous examples as the symmetric group on n letters. The idea of (linear) representation theory is that it is sometimes helpful to replace a group action on a set by a related action on a vector space.

DEFINITION 1.6. Suppose G acts on the set X . Write $\mathbf{C}X$ for the complex vector space of functions on X with values in \mathbf{C} . Define an action of G on $\mathbf{C}X$

by

$$(g \cdot f)(x) = f(g^{-1} \cdot x) \quad (g \in G, f \in CX, x \in X).$$

We call this action the *regular representation* of G on CX . We sometimes write the operators of this action as $\lambda(g)$: $\lambda(g)f = g \cdot f$.

The regular representation on functions on a G -space has the properties abstracted in the following definition.

DEFINITION 1.7. A (*linear*) *representation* of a group G is a vector space V endowed with an action of G by linear transformations. That is, it is a pair (π, V) , with V a vector space, and π a homomorphism from G into the group of automorphisms of V . We sometimes say that V *carries a representation of G* .

We will be interested almost exclusively in complex representations; that is, in the case when V is a vector space over \mathbb{C} . Some real representations will appear as auxiliary objects, however.

A reformulation along the lines of Definition (1.1)' is possible, and often helpful.

DEFINITION (1.7)' A representation of a group G on a vector space V is a map $G \times V \rightarrow V$, $(g, v) \rightarrow g \cdot v$, satisfying the following conditions:

- (a) for all g and h in G , and v in V , $g \cdot (h \cdot v) = (gh) \cdot v$;
- (b) if e is the identity element of G , then $e \cdot v = v$; and
- (c) if v and w are in V , and a and b are scalars, then $g \cdot (av + bw) = a(g \cdot v) + b(g \cdot w)$.

Just as in the case of general actions, it is this definition on which additional structure is almost easily imposed. For example, if G is a topological group, and V is a topological vector space, we say that the representation is *continuous* if the map from $G \times V$ to V is continuous.

The analogue for representations of transitive actions (Definition 1.2) is a little subtle.

DEFINITION 1.8. Suppose (π, V) is a representation of the group G . An *invariant subspace* of V is a linear subspace W of V , which is preserved by the action of G :

$$\pi(g \cdot w) \in W \quad (\text{all } g \in G, w \in W).$$

The representation is said to be *irreducible* if there are exactly two invariant subspaces.

A continuous representation is said to be *irreducible* if there are exactly two closed invariant subspaces. (This is not quite consistent with the nontopological definition.)

The subspaces $\{0\}$ and V are always invariant. To say that the representation is irreducible means that V is not zero, and that there are no other invariant subspaces.

The most obvious way to construct invariant subspaces is to start with a nonzero vector v , and look at the subspace $\langle G \cdot v \rangle$ generated by the orbit of v . These subspaces do not behave as well as orbits for actions, however. The difficulty is that a nonzero vector w in $\langle G \cdot v \rangle$ may in turn generate a strictly smaller

subspace. If V is infinite-dimensional, this process can go on forever; there may be no minimal invariant subspaces. Another aspect of the same problem is that even an irreducible representation is not characterized by the isotropy group of a single nonzero vector. There are no easy analogues of Lemmas 1.3 and 1.5 for linear representations. Nevertheless, our goal is to find some sort of analogues. A little more precisely, we have

ABSTRACT ABSTRACT HARMONIC ANALYSIS PROBLEM 1.9. I. For a reasonable representation (π, V) of a reasonable group G , show that (π, V) is (in some reasonable sense) a “direct sum” of irreducible representations of G .

II. For a reasonable group G , find all the reasonable irreducible representations of G .

Putting everything (symmetry and linearization, that is) together, we arrive at

CONCRETE ABSTRACT HARMONIC ANALYSIS PROBLEM 1.10. Suppose we have a mathematical problem P .

I. Translate P into a problem GXP about a space X on which a group G acts.

II. Translate GXP into a problem GVP about a vector space V of (something like) functions on X .

III. Decompose V as a representation of G , into irreducible representations V_i .

IV. Solve the problem for each irreducible representation V_i .

V. Combine these solutions into a solution of GVP .

VI. Translate back from GVP to P .

A shining example of a problem amenable to this process is the problem of finding the eigenvalues of the Laplace operator on the $(n-1)$ -dimensional sphere. There we can take X to be the sphere S^{n-1} ; G to be the orthogonal group $O(n)$; and V to be $L^2(S^{n-1})$. The result is the theory of spherical harmonics. We will give two more examples: one quite foolish but fairly simple; and the other quite technical, but illustrative of the success of the process in the context of reductive groups.

The first problem is that of finding the number N of p -element subsets of an n -element set S . We take X to be the set of such subsets, and V to be the space of all functions on X . The symmetric group G of permutations of S acts on V . We have $N = \dim V$. The irreducible representations of G are completely understood; they are parametrized by the partitions of n . Suppose that p is at most $\lfloor n/2 \rfloor$; the other case is similar. Then it turns out that V is a direct sum of exactly $p+1$ irreducible representations V_i ($i = 0, \dots, p$). Here V_i is the irreducible representation corresponding to the partition $(n-i, i)$. The dimensions of the irreducible representations are known; and

$$\dim V_i = [n!(n-2i+1)]/[i!(n-i)!(n-i+1)].$$

By induction on p , it is easy to show that

$$(\dim V_0) + \dots + (\dim V_p) = n!/p!(n-p)!.$$

This is then the dimension of V , and hence the cardinality of X .

The next example is Matsushima's formula; details about it may be found in [3] and [14]. Suppose M is a connected, compact, locally symmetric Riemannian manifold. ("Locally symmetric" means that the map sending v to $-v$ on the tangent space at each point p exponentiates to an isometry θ_p on some neighborhood of p . A Riemann surface always has this property, for example.) Consider the problem of finding the deRham cohomology of M . There is no group immediately evident in this problem; M may admit no isometries. But let M^\sim be the universal cover of M . The local isometries θ_p extend to all of M^\sim ; and so there is a large group G of isometries of M^\sim . The fundamental group Γ of M , regarded as the group of deck transformations of M^\sim , is a discrete subgroup of G . Set $X = G/\Gamma$, a compact homogeneous space for G . As a representation, we take $V = L^2(X)$. It turns out that V is a Hilbert space direct sum of a countable number of irreducible representations V_i ($i = 1, 2, \dots$).

To any nice continuous representation of G on W , it is possible to attach a collection of vector spaces $H^j(G, W)$, indexed by nonnegative integers j . These are called the *continuous cohomology of G with coefficients in W* . We have

$$H^0(G, W) = \{w \in W \mid g \cdot w = w, \text{ all } g \in G\};$$

and the higher H^j are defined as derived functors. It turns out that

$$H^j(M, \mathbb{C}) \cong H^j(G, V) \cong \bigoplus H^j(G, V_i);$$

this is Matsushima's formula. Finally, the groups $H^j(G, W)$ can be explicitly determined for any irreducible representation W occurring among the V_i .

The only part of this program which is not effective is the explicit determination of the V_i . Nevertheless, our explicit knowledge of the continuous cohomology of all possible V_i places rather strong *a priori* restrictions on what the cohomology of M can be like.

With applications such as this in mind, this paper will consider the following question: what can a reasonable representation of a reductive Lie group look like? (This question has by no means been answered completely.) §2 discusses what "reasonable" means, and indicates the rough shape of the answer suggested by the Kirillov-Kostant method of coadjoint orbits.

§3 shows how the Jordan decomposition of matrices separates the orbit method into three parts: hyperbolic, elliptic, and nilpotent. A little more precisely, it suggests that there ought to be three fundamental constructions of reasonable representations, and that all reasonable representations are obtained by applying them in succession.

The next three sections examine these three constructions. The hyperbolic step (giving continuous families of representations) is implemented by parabolic induction, developed in the 1950's by Gelfand-Naimark and Harish-Chandra. What is involved is fairly easy real analysis; the representation spaces are L^2 function spaces.

The elliptic step (giving discrete families of representations) was first seriously developed by Harish-Chandra in the 1960's, in his theory of square-integrable representations. It has been greatly extended in the past ten years, by Zuckerman and others. The methods are (in spirit, at least) complex-analytic; representations appear on (something like) Dolbeault cohomology spaces.

The nilpotent step is expected to give only a finite number of representations for each group. It has not yet been systematically developed beyond a very primitive form, and there is no general construction of the representations; but already it is possible to guess something about the appearance of the representations it will eventually produce.

I would like to thank Bert Kostant for teaching me about coadjoint orbits (or at least permitting me to drink from his firehose). Michel Duflo has explained a number of technical points.

2. Unitary representations and coadjoint orbits. Here is the class of "reasonable" representations we will consider in connection with Problem 1.9.

DEFINITION 2.1. Suppose G is a topological group. A representation (π, V) of G is called *unitary* if

- (a) π is continuous (defined after Definition (1.7)'),
- (b) V is a complex Hilbert space, and
- (c) for each element g of G , the operator $\pi(g)$ is unitary.

The set of equivalence classes of irreducible unitary representations of G is called the *unitary dual* of G , and written \hat{G} .

Even unitary representations need not decompose as (Hilbert space) direct sums of irreducible representations. An example is the regular representation of \mathbf{R} on $L^2(\mathbf{R})$ (Definition 1.6). In this example, the Fourier transform decomposes the space as a kind of continuous direct sum of irreducible representations. The "summands" are the spaces $V_t =$ multiples of the function χ_t , where $\chi_t(x) = e^{itx}$. (V_t is not a subspace of V . It is a representation of \mathbf{R} , however; it is a space of functions, and it is preserved by the regular representation.) This is like a direct sum in the sense that every vector v in $L^2(\mathbf{R})$ is written as a continuous combination $v = \int a_t \chi_t dt$. We summarize the basic L^2 properties of the Fourier transform by writing $V = \int_{\oplus} V_t dt$ and saying that V is the *direct integral* of the representations V_t .

We leave to the reader's imagination the general definition of the direct integral (of a family (π_x, V_x) of unitary representations, parametrized by the points of a measure space X). When it is properly formulated, we get

THEOREM 2.2 (SEE [10]). *Suppose π is a unitary representation of a group G on a separable Hilbert space V . Then (π, V) is equivalent to a direct integral of irreducible unitary representations. If G is a type I separable locally compact group, then this decomposition is unique.*

Type I is another term we prefer to leave undefined. Discrete groups are type I only if they are nearly abelian. All of your favorite Lie groups (for example,

the reductive ones as defined in §3 below) are type I. For part I of Problem 1.9, we have therefore found a reasonable class of groups to go with our reasonable class of representations.

The method of coadjoint orbits seeks to describe irreducible unitary representations of Lie groups in more or less geometric terms. A complete introduction to it may be found in [7, 9], and [5]. We will give an outline, concentrating on the goals of the method rather than on the means by which they are achieved. This requires some notation.

Suppose G is a Lie group. We will always assume that G has finitely many connected components. Put

$$\begin{aligned} G_0 &= \text{identity component of } G, \\ \mathfrak{g} &= \text{Lie algebra of } G, \\ \mathfrak{g}^* &= \text{vector space of real-valued linear functionals on } \mathfrak{g}. \end{aligned} \tag{2.3}$$

We regard \mathfrak{g} as the tangent space of G at the identity element; then \mathfrak{g}^* is the cotangent space at the identity.

Suppose G acts smoothly on a manifold X , and x is in X . Recall that $G(x)$ is the isotropy group at x (Definition 1.2); write $\mathfrak{g}(x)$ for its Lie algebra. The action mapping restricts to a smooth map

$$G \times \{x\} \rightarrow X, \quad g \rightarrow g \cdot x. \tag{2.4a}$$

The differential of this map at the identity is a map from \mathfrak{g} to $T_x(X)$ (the tangent space at X); its kernel is precisely $\mathfrak{g}(x)$. We get

$$\mathfrak{g}/\mathfrak{g}(x) \hookrightarrow T_x(X). \tag{2.4b}$$

If X is a homogeneous space, this is an isomorphism.

Fix now g in G , and consider the other restriction

$$\lambda_g: \{g\} \times X \rightarrow X, \quad x \rightarrow g \cdot x. \tag{2.5a}$$

The differential of this map is an isomorphism from $T_x(X)$ to $T_{g \cdot x}(X)$. In particular, we get the *isotropy representation*

$$(\tau_x, T_x(X)) \tag{2.5b}$$

of $G(x)$, defined by

$$\tau_x(g) = d\lambda_g(x) \quad (g \in G(x)). \tag{2.5c}$$

This is a finite-dimensional real representation.

G acts on itself by conjugation:

$$g \cdot x = gxg^{-1}. \tag{2.6a}$$

The isotropy group of the point e in G (the identity) is all of G . The isotropy representation (of G on the tangent space \mathfrak{g} of G at (e)) is called the *adjoint representation*, and written Ad :

$$\text{Ad}(g): \mathfrak{g} \rightarrow \mathfrak{g}. \tag{2.6b}$$

From any representation π on a vector space V , we can construct the *contragredient representation* π^* on the space V^* of linear functionals on V . We simply regard π as a group action, associate to it the regular representation of G on functions on V , and restrict to the linear functionals. Explicitly,

$$(\pi^*(g)\xi)(v) = \xi(\pi(g^{-1})v) \quad (2.7)$$

(for ξ in V^* , g in G , and v in V). The contragredient of the adjoint representation is called the *coadjoint representation*, and denoted

$$(\text{Ad}^*, \mathfrak{g}^*). \quad (2.8)$$

We will be interested in the coadjoint representation more as a group action than as a representation. That is, we will consider not invariant subspaces of \mathfrak{g}^* , but orbits. We will therefore speak of the *coadjoint action* and *coadjoint orbits*.

Here at last is the beginning of the Kirillov-Kostant orbit method.

PHILOSOPHY OF COADJOINT ORBITS (FIRST APPROXIMATION) 2.9. If G is a Lie group, the set \hat{G} of equivalence classes of irreducible unitary representations of G is related to the set of orbits of G on \mathfrak{g}^* .

This is a startling idea on its face; the two sets in question appear to be completely unrelated. Mathematical experience might lead one to suspect that there is a trick here—that, properly understood, irreducible unitary representations and coadjoint orbits will turn out to be the same thing almost by definition. There may be such a trick, but it has eluded group representers for more than twenty years now. I prefer to believe that there is real magic.

To understand this philosophy better, we need to make it more precise. The case of abelian groups is an example which helps. Notice first that a unitary operator on a one-dimensional complex Hilbert space is just multiplication by a scalar of absolute value 1. Write \mathbf{T} for the circle group. Then we have just seen that one-dimensional unitary representations are the same as homomorphisms into \mathbf{T} .

LEMMA 2.10. *Suppose H is an abelian group. Then any irreducible unitary representation has dimension one. If H is a connected abelian Lie group, then any such homomorphism χ is determined by its differential $d\chi$. If we identify $\text{Lie}(\mathbf{T})$ with $i\mathbf{R}$, then $d\chi$ corresponds in turn to a linear functional f on \mathfrak{g} . This gives an inclusion*

- (a) $\hat{G} \subset \mathfrak{g}^*$, $\chi \rightarrow f$, defined by
- (b) $\chi(\exp X) = \text{multiplication by } e^{if(X)}$.

Conversely, suppose f is any linear functional on H . Then f is identified with a character χ by (b), if and only if f maps the kernel of the exponential map into $2\pi\mathbf{Z}$:

- (c) $f(X) \in 2\pi\mathbf{Z}$, all $X \in \ker(\exp)$.

When H is abelian, an element $f \in \mathfrak{h}^*$ satisfying condition (c) is said to be *integral*. If H is simply connected, the condition is trivially satisfied for all f . If H is a torus, the kernel of the exponential map is a lattice in \mathfrak{h} ; the integral weights are just 2π times the dual lattice in \mathfrak{h}^* .

Lemma 2.10 suggests that the philosophy of coadjoint orbits ought to involve only orbits satisfying some kind of integrality condition. Here is a general formulation.

DEFINITION 2.11. Suppose G is a Lie group, and $f \in \mathfrak{g}^*$ (notation (2.3)). Recall that the isotropy group of the coadjoint action at f is written $G(f)$ (Definition 1.2), and that its identity component is $G(f)_0$. We say that f is *integral* if either of the following equivalent conditions is satisfied.

- (a) There is a unitary character $\pi_0(f)$ of $G_0(f)$, with differential if .
- (b) There is a finite-dimensional irreducible unitary representation $\pi(f)$ of $G(f)$, with differential $[d\pi(f)](X) = if(X) \cdot \text{Id}$ ($X \in \mathfrak{g}$).

We say that the pair $(f, \pi(f))$ is an *integral datum* for G . We can now formulate a better version of the orbit method.

PHILOSOPHY OF COADJOINT ORBITS (SECOND APPROXIMATION) 2.12. If G is a Lie group, then \hat{G} is in one-to-one correspondence with the set of integral data for G (Definition 2.11). In particular, \hat{G} is in finite-to-one correspondence with the set of integral coadjoint orbits for G . Lemma 2.10 says that this is true if G is connected and abelian. It is rather easy to deduce that it is also true when the identity component of G is abelian. In [6], it is proved to be true for G simply connected nilpotent; it follows in the case when the identity component of G is nilpotent.

The next obvious class of groups to consider is solvable groups. There are solvable Lie groups which are not type I (cf. Theorem 2.2). They do not satisfy Philosophy 2.12. This is almost good, however, because irreducible unitary representations are not the most important tools for harmonic analysis in the non-type I case. Auslander and Kostant showed in [2] that Philosophy 2.12 applies to simply connected type I solvable groups. As in the nilpotent case, we would like to drop the hypothesis that G be simply connected. There are formal problems with this, however. Although the integrality condition in Definition 2.11 is very pretty, it now turns out not to be quite right. The adjustments needed are rather delicate, but they involve some important ideas; so we will outline them.

Suppose V is a finite-dimensional real vector space. Suppose ω is a nondegenerate symplectic form on V ; that is, ω is a skew-symmetric bilinear form on V , with radical zero. Alternatively, one can think of ω as a 2-form on V . (Nondegeneracy in this interpretation means that for any nonzero v in V , there is a w such that $\omega(v \wedge w)$ is nonzero.) The *symplectic group* of V is

$$\text{Sp}(V) = \{g \in \text{GL}(V) | \omega(gv, gw) = \omega(v, w), \text{ all } v, w\}. \quad (2.13a)$$

(We may also write $\text{Sp}(\omega)$, or $\text{Sp}(V, \omega)$.) Its Lie algebra is

$$\mathfrak{sp}(V) = \{X \in \mathfrak{gl}(V) | \omega(Xv, w) + \omega(v, Xw) = 0\}. \quad (2.13b)$$

We want to define a two-fold cover $\text{Mp}(V)$ of $\text{Sp}(V)$. To do so properly would take a little too long. It is not hard to characterize the cover, however. If V is

zero, then $\mathrm{Sp}(V)$ is trivial; we define

$$\mathrm{Mp}(0) = \mathbf{Z}/2\mathbf{Z}. \quad (2.14)$$

If V is nonzero, we define $\mathrm{Mp}(V)$ to be the unique (connected) two-fold cover of $\mathrm{Sp}(V)$. In all cases, we get a short exact sequence

$$\{1, \varepsilon\} \rightarrow \mathrm{Mp}(V) \rightarrow \mathrm{Sp}(V). \quad (2.15)$$

$\mathrm{Mp}(V)$ is called the *metaplectic group*.

Suppose M is a manifold. A *symplectic structure* on M is a closed 2-form ω on M , which is nondegenerate as a bilinear form on each tangent space $T_m(M)$. If we are given such a structure, we say that M is a *symplectic manifold*.

THEOREM 2.16. *Suppose G is a Lie group, and F is a coadjoint orbit. Then F has a natural, G -invariant symplectic structure ω_F , defined as follows. Fix $f \in F$, and write ω_f for the symplectic form on the tangent space $T_f(F)$. Identify this tangent space with $\mathfrak{g}/\mathfrak{g}(f)$ (cf. (2.4b)). Given tangent vectors x and y , choose representatives X and Y in \mathfrak{g} . Then $\omega_f(x, y) = f([X, Y])$. The isotropy representation τ_f of $G(f)$ (cf. (2.5)) preserves ω_f .*

That the formula given defines a symplectic form on $T_f(F)$ is almost trivial. That ω_F is a closed 2-form on F is only slightly harder; it comes down to the Jacobi identity in \mathfrak{g} .

DEFINITION 2.17. Suppose G is a Lie group, and $f \in \mathfrak{g}^*$. By Lemma 2.16, there is a homomorphism $\tau_f: G(f) \rightarrow \mathrm{Sp}(\omega_f)$. The *metaplectic cover* of $G(f)$ is the pullback via τ of the metaplectic cover of $\mathrm{Sp}(\omega_f)$ (cf. (2.15)). It is denoted $G(f)^{\mathrm{mp}}$. We have

$$1 \rightarrow \{1, \varepsilon\} \rightarrow G(f)^{\mathrm{mp}} \rightarrow G(f) \rightarrow 1.$$

A representation $\pi(f)^{\mathrm{mp}}$ of $G(f)^{\mathrm{mp}}$ is called *genuine* if

(1) The nontrivial element ε of the kernel of the covering map acts by -1 in $\pi(f)^{\mathrm{mp}}$.

It is called *admissible* if it is genuine, and its differential satisfies

(2) $d\pi(f)^{\mathrm{mp}}(X) = if(X) \cdot \mathrm{Id}$.

In this case, the pair $(f, \pi(f)^{\mathrm{mp}})$ is called an *admissible datum* for G . The element f (or the orbit $G \cdot f$) is called *admissible* if there is an admissible representation of $G(f)^{\mathrm{mp}}$.

With this definition, Duflo has shown that all questions of covering groups fit together properly; irreducible unitary representations of type I solvable Lie groups are parametrized by G -conjugacy classes of admissible data. This suggests another approximation to the orbit method, which the reader can easily formulate.

When this philosophy is applied to semisimple groups (to be defined in §3 below), several things go wrong. The first problem appears when G is $\mathrm{SO}(3)$, the rotation group in three dimensions. In this case, the trivial representation of G is attached both to the orbit $\{0\}$, and to the smallest nonzero admissible

orbit. We can no longer expect to have a bijection between representations and admissible data.

A more fundamental problem appears when G is $\mathrm{SL}(2, \mathbf{R})$. There is a family of irreducible unitary representations called the *complementary series*, parametrized by the open unit interval $(0, 1)$. Except for the one parametrized by $\frac{1}{2}$, these representations appear not to be associated to any coadjoint orbit. The correspondence from orbits to representations is therefore not surjective.

Again for $\mathrm{SL}(2, \mathbf{R})$, there is a representation which appears to be attached to a union of two orbits, and not to either one individually. (It is the spherical principal series with normalized parameter zero.) The orbits in question are not closed, and their closures have a nonempty intersection.

Other minor problems along these general lines appear for more complicated semisimple groups: the representations attached to admissible data can be reducible, or zero, or they can even fail to exist. But philosophies are not as susceptible to counterexample as theorems, or even conjectures; and something like this at least survives.

PHILOSOPHY OF COADJOINT ORBITS (THIRD APPROXIMATION) 2.18. Suppose G is a type I Lie group. Attached to a finite set of admissible data for G (Definition 2.17), and some boundary conditions along the closures of the corresponding orbits, there is a unitary representation of G . All nice irreducible unitary representations arise in this way.

For the rest of this paper, we will be concerned with implementing the first part of this philosophy: attaching Hilbert spaces and operators to manifolds and group actions. We ignore completely two fascinating problems associated with the second part: giving an *a priori* definition of “nice,” and attaching orbits to (nice) representations.

3. Coadjoint orbits for reductive groups. The first reductive group to understand (even before reductive is defined) is $\mathrm{GL}(n, \mathbf{R})$, the group of n by n invertible real matrices. Its Lie algebra is

$$\mathfrak{gl}(n, \mathbf{R}) = \text{all } n \text{ by } n \text{ real matrices.} \quad (3.1a)$$

The adjoint action is by conjugation of matrices:

$$\mathrm{Ad}(g)(X) = gXg^{-1}. \quad (3.1b)$$

The Lie algebra carries a nondegenerate symmetric bilinear form, called the *trace form*, and denoted $\langle \cdot, \cdot \rangle$:

$$\langle X, Y \rangle = \mathrm{tr} \, XY. \quad (3.1c)$$

It is preserved by the adjoint action. The trace form defines an identification of $(\mathfrak{gl}(n, \mathbf{R}))^*$ with all n by n matrices. The linear functional f corresponds to the matrix $X(f)$ defined by

$$f(Y) = \langle X(f), Y \rangle. \quad (3.1d)$$

Because of the invariance of $\langle \cdot, \cdot \rangle$ under Ad , this identification sends coadjoint orbits to adjoint orbits. That is,

coadjoint orbits for $\text{GL}(n, \mathbf{R})$ are in one-to-one correspondence (3.2)
with conjugacy classes of n by n real matrices.

We recall now some elementary facts about conjugacy classes of matrices.

DEFINITION 3.3. Suppose X is an n by n real matrix. We say that X is *nilpotent* if X^k is zero for some k ; or, equivalently, if all the (complex) eigenvalues of X are zero. We say X is *semisimple* if X is diagonalizable over \mathbf{C} . It is *elliptic* if it is semisimple, and all the eigenvalues are purely imaginary. It is *hyperbolic* if it is semisimple and all the eigenvalues are real; or, equivalently, if it is diagonalizable over \mathbf{R} .

PROPOSITION 3.4 (JORDAN DECOMPOSITION). *Suppose X is a real n by n matrix. Then there are unique matrices X_h, X_e , and X_n , with the following properties:*

- (a) $X = X_h + X_e + X_n$;
- (b) X_h is hyperbolic, X_e is elliptic, and X_n is nilpotent; and
- (c) X_h, X_e , and X_n all commute with each other.

They have in addition the following property:

- (d) *any matrix commuting with X commutes also with X_h, X_e , and X_n .*

The usual Jordan decomposition, into semisimple and nilpotent parts, has semisimple part

$$X_s = X_h + X_e. \quad (3.5)$$

The *Cartan involution* for $\text{GL}(n, \mathbf{R})$ is the automorphism θ defined by

$$\theta g = {}^t g^{-1}, \quad (3.6a)$$

for g in $\text{GL}(n, \mathbf{R})$. Its differential, also denoted by θ , is the automorphism

$$\theta X = -{}^t X \quad (3.6b)$$

of $\mathfrak{gl}(n, \mathbf{R})$. (Since both group and algebra consist of matrices, the notation is inconsistent.) On the Lie algebra, the $+1$ eigenspace of θ is the Lie algebra of skew-symmetric matrices. It consists of elliptic elements, and the trace form is negative definite there. The -1 eigenspace consists of symmetric matrices, all of which are hyperbolic; the trace form is positive definite there.

At this point, it is convenient to introduce general reductive groups. The definition used here is borrowed from [8].

DEFINITION 3.7. A Lie group G (having finitely many components) is called *reductive* if there is a homomorphism $\eta: G \rightarrow \text{GL}(n, \mathbf{R})$ with the following properties:

- (1) the kernel of η is finite;
- (2) the image of η is θ -stable.

G is called *semisimple* if it is reductive, and the center of G_0 is finite.

We write θ for the unique lifting of θ to G which is trivial on the kernel of η , and call it the *Cartan involution of G* . The differential of η identifies \mathfrak{g} with a Lie algebra of matrices. Elements of \mathfrak{g} are called *semisimple*, *nilpotent*, *hyperbolic*, or *elliptic* when the corresponding matrices are. Put

$$\begin{aligned} K &= \text{fixed points of } \theta \text{ on } G, \\ \mathfrak{t} &= \text{Lie}(K) = \text{fixed points of } \theta \text{ on } \mathfrak{g}, \\ \mathfrak{s} &= -1 \text{ eigenspace of } \theta \text{ on } \mathfrak{g}. \end{aligned}$$

We get the *Cartan decomposition* $\mathfrak{g} = \mathfrak{t} + \mathfrak{s}$. By the remarks after (3.6), the trace form $\langle \cdot, \cdot \rangle$ is positive definite on \mathfrak{s} and negative definite on \mathfrak{t} . By the Cartan decomposition, $\langle \cdot, \cdot \rangle$ is therefore nondegenerate. We use it as in (3.1d) to identify \mathfrak{g}^* with \mathfrak{g} : the linear functional f on \mathfrak{g} corresponds to the element $X(f)$ satisfying $f(Y) = \langle X(f), Y \rangle$, for all Y in \mathfrak{g} .

Here are some important structural facts. For $\text{GL}(n, \mathbf{R})$, they amount to Proposition 3.4, and the fact that every real elliptic matrix is conjugate to a skew-symmetric one.

PROPOSITION 3.8. *Suppose G is a real reductive group, and X is in \mathfrak{g} .*

(a) *The components X_h, X_e , and X_n of the Jordan decomposition of X all lie in \mathfrak{g} .*

(b) *If X is hyperbolic, it is conjugate under $\text{Ad}(G)$ to an element of \mathfrak{s} .*

(c) *If X is elliptic, it is conjugate under $\text{Ad}(G)$ to an element of \mathfrak{t} .*

DEFINITION 3.9. Suppose G is a reductive group, and $f \in \mathfrak{g}^*$. Write $X(f)$ for the corresponding element of \mathfrak{g} (Definition 3.7), and

$$X(f) = X(f)_h + X(f)_e + X(f)_n$$

for its Jordan decomposition (Proposition 3.4). The corresponding linear functionals on \mathfrak{g} are written f_h, f_e , and f_n ; and $f = f_h + f_e + f_n$ is the *Jordan decomposition* of f . We call f *hyperbolic*, *nilpotent*, etc. if $X(f)$ is. The *semisimple part* of f is $f_s = f_h + f_e$.

Here is how the Jordan decomposition is to be used to organize the problem of associating representations to orbits. Suppose we are given f . We will use constantly the fact that

$$G(f) = \text{centralizer of } X(f) \text{ in } G, \quad (3.10a)$$

$$\mathfrak{g}(f) = \{Y \in \mathfrak{g} \mid [X(f), Y] = 0\}. \quad (3.10b)$$

Proposition 3.8 allows us to replace f by a conjugate, and get

$$\theta f_h = -f_h, \quad \theta f_e = f_e. \quad (3.11a)$$

It follows that the isotropy groups

$$G(f_h), \quad G(f_e), \quad \text{and} \quad G(f_s) = G(f_h) \cap G(f_e) \quad (3.11b)$$

are all preserved by θ ; they are therefore reductive (via the restrictions of the map η used for G itself). The elements X_e and X_n commute with X_h , and

so belong to $\mathfrak{g}(f_h)$. We can therefore identify f_e and f_n (by restriction) with elements of $\mathfrak{g}(f_h)^*$. We get a chain

$$G(f_h) \supset [G(f_h)](f_e) \supset [[G(f_h)](f_e)](f_n); \quad (3.11c)$$

these are the same groups as

$$G(f_h) \supset G(f_s) \supset G(f). \quad (3.11c)'$$

The datum we are given is (more or less) a representation of $G(f)$. From this, we propose to get a representation of G in three steps. First, we will get a representation of $G(f_s)$ (the nilpotent step); then a representation of $G(f_h)$ (the elliptic step); then a representation of G (the hyperbolic step). The terminology arises because (for example) the subgroup $G(f_s)$ (from which we begin in the elliptic step) is the isotropy group for the elliptic element f_e in $G(f_h)$ (to which we are going).

As was indicated in the introduction, the order of these steps exactly reverses the historical one; and in fact the general treatment of the first step is still in the future. Undaunted by a suspicion that nothing comes from nothing, however (and having utterly misspent our youths and childhoods), we will treat the last two steps in the next two sections. The last section will discuss prospects for the first step.

4. Parabolic induction and the hyperbolic step. All that we know so far of coadjoint orbits is that they are symplectic homogeneous spaces. (For semisimple groups, that is all there is to know: Kirillov, Kostant, and Souriau have shown independently that any such space is a finite covering of an orbit.) The orbit method asks us to build a representation out of an orbit $X = G \cdot f$. There is only one obvious representation in sight, namely the regular representation on X (Definition 1.6), or something closely related to it. Experimental evidence shows that this is too large: it is almost never irreducible, for example.

A more careful analysis would suggest building a bundle on X out of the admissible datum $(f, \pi(f)^{\text{mp}})$. The difficulty is that $\pi(f)^{\text{mp}}$ must first be untwisted into a representation of $G(f)$ (and not just of $G(f)^{\text{mp}}$). This seems to require something like tensoring it with the metaplectic representation of $\text{Mp}(\omega_f)$. We are left with an infinite-dimensional bundle on a space which was too large to begin with. This looks bad enough to be promising, but no progress has been made in this direction.

If the symplectic structure does not suffice to produce a representation, it is reasonable to ask what additional structure would help. A useful way to phrase that question is this: what more complicated objects happen to be symplectic manifolds in a natural way?

The first answer is cotangent bundles. Suppose Y is any manifold. Then T^*Y carries a natural symplectic structure. If G acts on Y , then it acts on T^*Y , preserving this structure. To the symplectic manifold T^*Y , one can associate the unitary representation of G on $L^2(Y)$. More precisely, we should consider

the representation on square-integrable sections of the half-density bundle on Y . This makes sense even if Y has no invariant measure.

More generally, suppose \mathcal{L} is a Hermitian line bundle on Y . Then R. Urwin has shown that there is a *connection bundle* $\mathcal{C}_{\mathcal{L}}$ over Y , such that a selfadjoint connection on \mathcal{L} is precisely a section of $\mathcal{C}_{\mathcal{L}}$. The space $\mathcal{C}_{\mathcal{L}}$ has a symplectic structure; and if \mathcal{L} is a homogeneous line bundle, then G acts on $\mathcal{C}_{\mathcal{L}}$. To the symplectic G -space $\mathcal{C}_{\mathcal{L}}$, we associate the representation of G on square-integrable sections of \mathcal{L} (twisted by half densities).

The first technique for attaching a representation to an orbit X is therefore to try to realize X as (roughly speaking) the cotangent bundle of some homogeneous space Y ; or (more precisely) as the connection bundle for a homogeneous line bundle on Y .

In the case of reductive groups, we can apply this technique to hyperbolic elements. Fix f_h in \mathfrak{g}^* hyperbolic, and write X_h for the corresponding Lie algebra element (Definition 3.7). We have an eigenspace decomposition

$$\mathfrak{g} = \sum_{r \in \mathbb{R}} \mathfrak{g}^r. \quad (4.1a)$$

Here

$$\mathfrak{g}^r = \{Y \in \mathfrak{g} \mid [X_h, Y] = rY\}. \quad (4.1b)$$

By (3.10), $G(f_h)$ preserves this decomposition, and

$$\mathfrak{g}^0 = \mathfrak{g}(f_h). \quad (4.1c)$$

The Jacobi identity shows that

$$[\mathfrak{g}^r, \mathfrak{g}^s] \subset \mathfrak{g}^{r+s}. \quad (4.1d)$$

The $\text{ad}(X_h)$ -invariance of the trace form $\langle \cdot, \cdot \rangle$ shows that

$$\langle \mathfrak{g}^r, \mathfrak{g}^s \rangle = 0 \quad \text{if } r + s \neq 0. \quad (4.1e)$$

Define

$$\mathfrak{n}_h = \sum_{r > 0} \mathfrak{g}^r. \quad (4.2a)$$

By (4.1), \mathfrak{n}_h is a nilpotent subalgebra of \mathfrak{g} , normalized by $G(f_h)$. Define

$$N_h = \exp(\mathfrak{n}_h), \quad (4.2b)$$

$$P_h = G(f_h)N_h. \quad (4.2c)$$

The group P_h is what is called a *parabolic subgroup* of G . We are trying to make the space $G/G(f_h)$ look like a certain kind of bundle over a smaller homogeneous space. The smaller space will be G/P_h .

Here is an outline of the hyperbolic step of the orbit method. Recall that the preceding steps (yet to be discussed) are to have given us (roughly) a unitary representation π_h of $G(f_h)$. Extend this representation to all of P_h by making it trivial on N_h . Form the induced Hermitian bundle \mathcal{V}_h on G/P_h . The representation of G that we want is that on the space of square-integrable global sections

of \mathcal{V}_h . More precisely, we want sections of \mathcal{V}_n twisted by the half-density bundle on G/P_h .

The experts will note that this can be elaborated into an orbit-theoretic partial classification of unitary representations, in the spirit of Duflo's theorem in [4] for general algebraic Lie groups. A datum for this partial classification is a pair (f, π^{mp}) (up to conjugacy by G). Here f is hyperbolic; π^{mp} is an irreducible unitary genuine representation of $G(f)^{\text{mp}}$ (Definition 2.17); and the imaginary part of the infinitesimal character of π^{mp} is $i \cdot f$.

5. Cohomological induction and the elliptic step. Having done what we can with cotangent bundles, we ask again: what other objects are also symplectic manifolds? The next answer we consider is: Kähler manifolds. These are manifolds which are both complex and symplectic, in a compatible way.

More precisely, suppose V is a real vector space. Recall that a *complex structure* on V is a map

$$J: V \rightarrow V \quad (5.1a)$$

such that

$$J^2 = -\text{Id}. \quad (5.1b)$$

Here is another formulation. Write

$$V_{\mathbb{C}} = V \otimes_{\mathbb{R}} \mathbb{C} = \{v + iw|v, w \in V\}, \quad (5.2a)$$

the *complexification* on V . $V_{\mathbb{C}}$ is a complex vector space. *Complex conjugation* on V is the conjugate linear automorphism σ defined by

$$\sigma(v + iw) = v - iw. \quad (5.2b)$$

Giving a complex structure on V is equivalent to giving a complex subspace

$$V^{0,1} \subset V_{\mathbb{C}}, \quad (5.3a)$$

with the property that

$$V_{\mathbb{C}} = \sigma(V^{0,1}) \oplus V^{0,1} = V^{1,0} \oplus V^{0,1}. \quad (5.3b)$$

We call $V^{0,1}$ the *anti-holomorphic subspace*. (The equivalence sends J to the $-i$ eigenspace of J on $V_{\mathbb{C}}$.)

DEFINITION 5.4. A *Kähler structure* on a real vector space V consists of

(1) a complex structure J , and

(2) a symplectic structure ω

(cf. (5.1) and (2.13)). These are required to satisfy

(a) $\omega(Jv, w) = -\omega(v, Jw)$.

An equivalent condition is

(a)' $\omega(Jv, Jw) = \omega(v, w)$.

If the complex structure is considered to be defined by the subspace $V^{0,1}$, then the requirement is equivalent to

(a)'' $\omega_{\mathbb{C}}(v, w) = 0$, all $v, w \in V^{0,1}$.

Here $\omega_{\mathbb{C}}$ denotes the complex-linear extension of ω to $V_{\mathbb{C}}$.

An equivalent formulation is this: a Kähler structure on V consists of

- (1) a complex structure J , and
- (2) a nondegenerate Hermitian form on the (complex) vector space V .

(Recall that a Hermitian form is a complex-valued sesquilinear form on V , satisfying $\langle v, w \rangle = \overline{\langle w, v \rangle}$.) The two definitions are related by

$$\omega(v, w) = \operatorname{Im} \langle v, w \rangle, \quad \langle v, w \rangle = \omega(Jv, w) + i\omega(v, w).$$

We call $\langle \cdot, \cdot \rangle$ the *Kähler form* on V . Its signature (p, q) is called the *signature* of the Kähler structure.

Finally, we can think of a Kähler structure as consisting of

- (1) a nondegenerate symmetric bilinear form B on V , and
- (2) a nondegenerate symplectic form W on V .

These are subject to the following condition: suppose the automorphism J of V is defined by

$$(a) \quad B(v, w) = \omega(Jv, w).$$

Then $J^2 = -\operatorname{Id}$.

DEFINITION 5.5. Suppose M is a manifold. A *Kähler structure* on M is a complex structure and a symplectic structure, which give a Kähler structure on each tangent space $T_m M$.

Thus a Kähler manifold is simultaneously complex, symplectic, and (possibly indefinite) Riemannian; and any two of these structures determine the third.

From the point of view of the orbit method, recall that what we seek is a smaller version of the space of functions (or sections of a line bundle) on M . A natural choice is the holomorphic functions. Just as the functions on a space Y have (morally) half as many degrees of freedom as those on T^*Y , so also the holomorphic functions on a complex manifold have half the freedom of the smooth ones.

We tread on thin analytic ice here, however. The absence of “bump functions” subjects holomorphic functions to global constraints which have no parallel in real analysis. To understand the problem, let us consider an example.

EXAMPLE 5.6. Take G to be $\operatorname{GL}(2n, \mathbf{R})$. Recall that \mathfrak{g}^* consists of $2n$ by $2n$ matrices, that is, of linear transformations of \mathbf{R}^{2n} . Fix an identification of \mathbf{R}^{2n} with \mathbf{C}^n , and let

$$(a) \quad f = \text{matrix of multiplication by } i.$$

If the identification is made properly,

$$(b) \quad f = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$$

This element is elliptic. A moment's thought shows that its centralizer looks like

$$(c) \quad G(f) \cong \operatorname{GL}(n, \mathbf{C}).$$

The orbit $F = G \cdot f$ consists of all matrices f' with square -1 , that is, of all complex structures on \mathbf{R}^{2n} .

We claim that F carries a Kähler structure, or, what amounts to the same thing, that it has a G -invariant complex structure giving (with ω_f) a Kähler structure on $\mathfrak{g}/\mathfrak{g}(f)$. To see that, we use the second interpretation of complex structures, as subspaces of $(\mathbf{R}^{2n})_{\mathbf{C}} = \mathbf{C}^{2n}$. The space F is now identified with a

subset of the (compact complex) Grassman manifold $\text{Gr}(n, 2n)$, of n -dimensional subspaces in \mathbf{C}^{2n} . The condition on the subspace imposed by (5.3b) is open, so F is open in $\text{Gr}(n, 2n)$. This gives an invariant complex structure; we omit the verification that F is Kähler. The complex dimension of F is n^2 ; the signature of the Kähler form is $((n^2 - n)/2, (n^2 + n)/2)$.

The homogeneous Hermitian line bundles on F are parametrized by unitary characters of the isotropy group $\text{GL}(n, \mathbf{C})$. They in turn are parametrized by \mathbf{Z} , by

$$\chi_m(g) = (\det(g)/|\det(g)|)^m.$$

Write \mathcal{L}_m for the bundle corresponding to χ_m . It turns out that the space of holomorphic sections \mathcal{L}_m is always finite-dimensional; the dimension is positive exactly when m is at most zero. The representations of G which arise on the holomorphic sections are the Cartan powers of the fundamental representation attached to the middle simple root; they are never unitary, except for the trivial representation when m is zero.

Part of the difficulty is that the complex manifold F has a large compact subvariety. Write B for the standard inner product on \mathbf{R}^{2n} , extended by complex linearity to \mathbf{C}^{2n} . Consider the following subspace of $\text{Gr}(n, 2n)$:

$$(d) F_K = \{V | B(V, V) = 0\}.$$

It is evidently an algebraic subvariety. The fact that B is definite on \mathbf{R}^{2n} implies that F_K is contained in F . In fact it is the orbit of F under the orthogonal group:

$$(e) F_K = \text{O}(2n) \cdot f \cong \text{O}(2n)/\text{U}(n).$$

F_K is therefore a compact complex subvariety of F , of complex dimension $(n^2 - n)/2$.

A number of clues here suggest that instead of using holomorphic functions, one ought to consider the higher Dolbeault cohomology of F with coefficients in \mathcal{L}_m ,

$$(f) V^p(m) = H^{0,p}(F, \mathcal{L}_m).$$

(This suggestion was made in a slightly different context by Kostant and Langlands, about twenty years ago.) These spaces at least carry representations of G .

One good clue as to the nature of the representations is this. Put

$$(g) (V^p(m))_K = H^{0,p}(F_K, \mathcal{L}_m).$$

This finite-dimensional space carries a representation of K , which is explicitly computable by the Bott-Borel-Weil theorem. We find

$$(h) (V^p(m))_K \neq 0 \text{ iff either } p = 0 \text{ and } m \leq 0, \text{ or } p = (n^2 - n)/2 \text{ and } m \geq (n - 1).$$

The restriction map for cohomology gives a map from $V^p(m)$ to $(V^p(m))_K$. This suggests that $V^p(m)$ is most interesting under the conditions in (h) above.

We have already discarded the case of nonpositive m as uninteresting; so we are forced to consider the other case. The determinant of the action of $G(f)$ on the holomorphic cotangent space at zero is χ_{2n} , so one can imagine that there is a "half-density" shift by n involved somewhere.

What finally emerges (after some more serious labor) is this. Suppose m is greater than or equal to n , and p is $(n^2 - n)/2$. There there is a G -invariant dense subspace

$$(i) \mathcal{H}(m - n) \subset V^p(m),$$

which has a G -invariant inner product making it into a Hilbert space. For m greater than n , the resulting unitary representation is the one attached to the orbit $[(m - n)/2]F$.

This example is very helpful in answering the question of approximately how representations ought to be constructed from elliptic orbits. It is much less clear in detail: we brought an inner product into the picture at the end of the example, entirely out of nowhere. The alert reader may have noticed some difficulty earlier as well: the Dolbeault cohomology space arises as (for example) the cohomology of a complex of $(0, p)$ -forms. There is no easy reason for the differential to have closed range, so it is difficult even to define a topology on $V^p(m)$.

On compact manifolds, both problems are cured by Hodge theory: one finds harmonic representatives of everything, and does analysis with those. There are three difficulties with that idea in the present case. First, the Kähler form is indefinite, so the Laplacian is not elliptic. The harmonic forms are therefore not quite so nice. Second, the manifold is noncompact; so there are convergence questions to answer before one can integrate forms to get an inner product on cohomology. Third, the indefiniteness of the Kähler form makes the local inner product on $(0, p)$ -forms indefinite; so it is not clear that the inner product on cohomology is positive.

For the moment, these problems appear insuperable. (Schmid has solved them for the discrete series. Some inroads are made on the general case in [11].) For the purposes of representation theory, however, they may be regarded as completely resolved by Zuckerman's theory of cohomological parabolic induction (see [12]). That theory builds representations using a formal imitation of complex analysis on appropriate homogeneous spaces. It is Zuckerman's method which allows us to complete the elliptic step of the orbit method. To simplify the description, however, we will stay in the complex analysis setting.

Suppose now that G is reductive, and that f_e is an elliptic element of \mathfrak{g}^* . Write X_e for the corresponding Lie algebra element. It turns out that $\text{ad}(X_e)$ has purely imaginary eigenvalues, so we have an eigenspace decomposition

$$\mathfrak{g}_{\mathbb{C}} = \sum_{r \in \mathbb{R}} (\mathfrak{g}_{\mathbb{C}})^r. \quad (5.7a)$$

Here

$$(\mathfrak{g}_{\mathbb{C}})^r = \{Y \in \mathfrak{g} \mid [iX_e, Y] = rY\}. \quad (5.7b)$$

This is preserved by $G(f_e)$, and the zero weight space is $\mathfrak{g}(f_e)_{\mathbb{C}}$. The analogues of (4.1d) and (e) hold as well.

Define

$$u_e = \sum_{r > 0} (\mathfrak{g}_{\mathbb{C}})^r \quad (5.8a)$$

and

$$\mathfrak{q}_e = \mathfrak{g}(f_e)_{\mathbb{C}} + \mathfrak{u}_e. \quad (5.8b)$$

The subspace

$$\mathfrak{q}_e/\mathfrak{g}(f_e)_{\mathbb{C}} \subset [\mathfrak{g}/\mathfrak{g}(f_e)]_{\mathbb{C}} \quad (5.8c)$$

turns out to be the anti-holomorphic tangent space at f_e for a Kähler structure on the orbit

$$F_e = G \cdot f_e. \quad (5.9a)$$

Write

$$(s, r) = \text{signature of Kähler form on } F_e. \quad (5.9b)$$

Here is an outline of the elliptic step of the orbit method. (When this is part of the program described at (3.11), G is replaced by $G(f_h)$.) The unipotent step is supposed to have given us a genuine unitary representation π^{mp} of $G(f_e)^{\text{mp}}$. After a twist by the square root of the determinant of the action of $G(f_e)$ on the holomorphic cotangent space at f_e , we get precisely a unitary representation (π_e, V_e) of $G(f_e)$. This induces a holomorphic Hilbert bundle \mathcal{V}_e on F_e . The representation we want is on an appropriate dense subspace of the Dolbeault cohomology

$$H^{0,s}(F_e, \mathcal{V}_e). \quad (5.10)$$

As stated earlier, there is an existence theorem for these unitary representations.

“THEOREM” 5.11 [12]. *Suppose G is a reductive group, and f_e is an elliptic element in \mathfrak{g}^* . Use the notation of (5.7) – (5.9). Recall from Definition 2.17 the metaplectic cover $G(f_e)^{\text{mp}}$.*

(a) *There is a genuine character ρ_e of $G(f_e)^{\text{mp}}$, such that $[(\rho_e)(g)]^2 = \text{determinant of } \text{Ad}(g) \text{ on } \mu_e$, for g in $G(f_e)^{\text{mp}}$.*

Fix an irreducible unitary genuine representation π^{mp} of $G(f_e)^{\text{mp}}$. Assume that (1) the restriction of π^{mp} to the commutator subgroup is weakly unipotent [12, Definition 8.16]; and (2) π^{mp} has differential $(if_e)\text{Id}$ on the center of $\mathfrak{g}(f_e)$. Set $\pi_e = \pi^{\text{mp}} \otimes \rho_e$, and let \mathcal{V}_e be the induced holomorphic Hilbert bundle on F_e .

(b) *The Dolbeault cohomology of F_e with coefficients in \mathcal{V}_e vanishes except in degree s .*

(c) *There is a dense G -invariant subspace $V(f_e, \pi^{\text{mp}}) \subset H^s(F_e, \mathcal{V}_e)$, which carries a unitary representation of G .*

What the quotation marks mean is this. The result is probably true as stated, but current techniques prove only an algebraic analogue. (An honest unitary representation is given by the algebra; all that is lacking is the geometric realization.) The first of the two hypotheses can be regarded as a desired property of the orbit method for nilpotent orbits. The second is tied to the first hypothesis in Definition 2.17. One way that both hypotheses can be satisfied is if π^{mp} is a unitary character with differential f_e .

The representation V given by the theorem may be reducible or zero; but neither of these possibilities can arise if f_e is large enough.

6. The nilpotent step. What remains is to understand what representations of our reductive group G are associated to nilpotent coadjoint orbits. This is still largely mysterious. The nilpotent orbits are typically not cotangent bundles; except for the case of the point zero, they never admit invariant Kähler structures; and group representers have thought of no other answers to the question posed at the beginning of §4.

Nevertheless, the methods of the hyperbolic and elliptic steps have something to offer here. First, the element zero of \mathfrak{g}^* is both hyperbolic and elliptic (as well as nilpotent); either §4 or §5 says that the representations associated to it must be those trivial on G_0 . Here is a more refined version of the same idea.

LEMMA 6.1. *In the setting of equations (4.1) and (4.2), there are open orbits (E_1, \dots, E_t) of G on $T^*(G/P_h)$. Each E_i is a finite cover of a nilpotent coadjoint orbit F_i (as a symplectic homogeneous space).*

This lemma suggests the following requirement.

REQUIREMENT 6.2. In the setting of Lemma 6.1, suppose \mathcal{V} is a homogeneous Hermitian vector bundle on G/P_h , with an invariant flat connection. Then the unitary representation of G on L^2 sections of \mathcal{V} (twisted by half-densities) must be one of those associated to the nilpotent coadjoint orbits F_i .

A requirement along the same lines can be made using the elliptic step.

REQUIREMENT 6.3. Suppose G is a reductive group, and f_e is an elliptic element in \mathfrak{g}^* . Use the notation (5.7)–(5.9). Suppose π^{mp} is a genuine representation of $G(f_e)^{\text{mp}}$, trivial on the identity component. Set

$$\pi_e = \pi^{\text{mp}} \otimes \rho_e$$

(cf. Theorem 5.11), and let \mathcal{V}_e be the induced holomorphic Hilbert bundle on F_e . Then the unitary representation of G which is dense in $H^s(F_e, \mathcal{V}_e)$ must be one of those attached to nilpotent coadjoint orbits.

The nilpotent orbits to which this representation ought to be attached are those in the “associated cone”

$$\lim_{t \rightarrow 0^+} tF_e. \quad (6.4)$$

It can be shown that the representation is nonzero if and only if this limit cone has the same dimension as F_e . (The orbits in Requirement 6.2 may be obtained as the associated cone for F_h . They always have the same dimension as F_h .)

By arguments like this, one can collect evidence for proposed descriptions of the orbit method for nilpotent orbits. Additional help has come from primitive ideal theory and the theory of automorphic forms. One would like to apply knowledge about representations to these subjects. However, each of them has its own internal intuition, and it is possible to predict what the representation theory ought to say about them. These predictions can then be read as conjectures about representation theory—with luck, as descriptions of the representations attached to nilpotent orbits. The paper [1] is a landmark in this direction; an account of further developments will appear in [13].

REFERENCES

1. J. Arthur, *On some problems suggested by the trace formula*, Lie Group Representations, II (College Park, Md., 1982/1983), Lecture Notes in Math., vol. 1041, Springer-Verlag, Berlin-Heidelberg-New York, 1983.
2. L. Auslander and B. Kostant, *Polarization and unitary representations of solvable Lie groups*, Invent. Math. **14** (1971), 255–354.
3. A. Borel and N. Wallach, *Continuous cohomology, discrete subgroups, and representations of reductive groups*, Princeton Univ. Press, Princeton, N.J., 1980.
4. M. Duflot, *Théorie de Mackey pour les groupes de Lie algébriques*, Acta Math. **149** (1982), 153–213.
5. V. Guillemin and S. Sternberg, *Geometric asymptotics*, Mathematical Surveys, No. 14, Amer. Math. Soc., Providence, R.I., 1978.
6. A. Kirillov, *Unitary representations of nilpotent Lie groups*, Uspekhi Mat. Nauk. **17** (1962), 57–110.
7. A. Kirillov, *Elements of the theory of representations*, translated by E. Hewitt, Grundlehren der Mathematischen Wissenschaften, Band 220, Springer-Verlag, Berlin-Heidelberg-New York, 1976.
8. A. Knapp, *Representation theory of real semisimple groups: an overview based on examples*, Princeton Univ. Press, Princeton, N.J., 1986.
9. B. Kostant, *Quantization and unitary representations*, Lectures in Modern Analysis and Applications (C. Taam, ed.), Lecture Notes in Math., vol. 170, Springer-Verlag, Berlin-Heidelberg-New York, 1970.
10. G. Mackey, *Theory of unitary group representations*, Univ. of Chicago Press, Chicago, Ill., 1976.
11. J. Rawnsley, W. Schmid, and J. Wolf, *Singular unitary representations and indefinite harmonic theory*, J. Funct. Anal. **51** (1983), 1–114.
12. D. Vogan, *Unitarizability of certain series of representations*, Ann. of Math. **120** (1984), 141–187.
13. ———, *Unitary representations of reductive Lie groups*, Ann. of Math. Studies (to appear).
14. D. Vogan and G. Zuckerman, *Unitary representations with non-zero cohomology*, Compositio Math. **53** (1984), 51–90.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS
02139, USA

Physics and Geometry

EDWARD WITTEN

In many past epochs, problems arising from theoretical physics influenced the development of mathematics, or structures that first arose in mathematics entered in the development of physics. Famous twentieth-century examples would be the role of Riemannian geometry in facilitating the invention of general relativity or the influence of quantum mechanics in the development of functional analysis. The above-cited examples, however, involve innovations in physics that took place sixty or seventy years ago. In the last half century, mathematics and physics developed in very different directions, and interaction between the two disciplines played a less extensive role.

In part this happened because mathematics progressed into abstract realms seemingly unrelated to the humdrum world of the theoretical physicist. In part, it resulted from the way that physics developed. The two basic theories in twentieth-century physics are general relativity and quantum field theory. Their successes are in very different realms. General relativity—Einstein's theory of gravity—has its successful applications to large-scale astronomical phenomena, while quantum field theory is the framework within which physicists have been able to understand many properties of the elementary particles. General relativity was put in its final form by Einstein in 1915, while quantum field theory has been an open frontier since its formulation in the late 1920s. For half a century, the really fundamental advances in physics have mainly been developments in quantum field theory. In this period, it is quantum field theory which has been the central arena for possible interaction between mathematics and physics.

For some decades after the invention of quantum field theory, this theory was formulated in a rather technical and clumsy way, hard to work with even for physicists. It was not at all obvious that quantum field theory really existed as a sound mathematical theory. Most important, in the first few decades of quantum field theory, this subject did not give rise to very many interesting mathematical structures.

The outlook changed in the mid-1970s after nonabelian gauge theories emerged as the quantum field theories most relevant to physics. In the context of these

Research supported in part by National Science Foundation Grant PHY80-19754

theories, many significant physical problems lead to significant concepts in modern mathematics. For example, the study of magnetic monopoles and instantons involves the topological classification of vector bundles. The solution to the “U(1) problem” of quantum chromodynamics turned out to involve the Atiyah-Singer index theorem. The proper understanding of local and global “anomalies” involves fairly subtle properties of families of elliptic operators. Various other examples could be cited.

It certainly is charming to see “practical” applications of some seemingly abstruse mathematics. In some of the cases I have mentioned, the solution of the physics problem has actually required the uncovering of new mathematical theorems. All the same, the mutual interaction of mathematics and physics would remain rather limited, I believe, if it were only a question of quantum field theory. The applications of modern mathematics to quantum field theory are fascinating but relatively specialized; and the same can be said for the role that quantum field theory has so far played in stimulating mathematical innovations. It is in trying to go beyond the limitations of quantum field theory that physicists have really begun to meet mathematical frontiers.

The basic limitation of quantum field theory is that, as we noted earlier, it is only one of two fundamental theories in twentieth-century physics, the second being general relativity. Both of these theories play a role in describing the same natural world, so a more complete description of nature must encompass both of them. It has been rather clear, however, since the early days of quantum field theory that there are severe difficulties in trying to combine quantum field theory with general relativity. The formal attempt to quantize general relativity leads to nonsensical infinite formulas. In its early days, quantum field theory faced many difficulties, of which this was only one. As the other difficulties were overcome and quantum field theory emerged as an adequate framework for describing all of the natural forces except gravity, the inconsistency between general relativity and quantum field theory emerged clearly as *the* limitation of quantum field theory.

This problem is a theorists’ problem par excellence. Experiment provides little guide except for the bare fact that quantum field theory and general relativity both play a role in the description of natural law. Unfortunately, gravitational effects are unmeasurably small in all feasible experiments in which quantum field theory plays an observable role—and vice versa. All the same, the inconsistency between the two central theories in physics is clearly an important problem on the logical plane. Indeed, the history of physics gives many examples showing how important such problems are. For example, general relativity was invented in Einstein’s effort to resolve an inconsistency between two leading theories of that time, namely, special relativity and Newtonian gravity. Quantum field theory was similarly born in an attempt to reconcile nonrelativistic quantum mechanics with special relativity.

In the discovery of general relativity, the logical framework came first. Einstein first thought through the physical principles which the new theory should

embody, then found in Riemannian geometry the correct mathematical framework, and finally formulated the theory. The development of quantum mechanics and quantum field theory was quite different. There was no a priori conceptual insight; experimental clues played an extensive role. As I have indicated, experiment is not likely to provide detailed guidance about the reconciliation of general relativity with quantum field theory. One might therefore believe that the only hope is to emulate the history of general relativity, inventing by sheer thought a new mathematical framework which will generalize Riemannian geometry and will be capable of encompassing quantum field theory. Many ambitious theoretical physicists have aspired to do such a thing, but little has come of such efforts.

Progress seems to have come, instead, in a rather different way. In the course of attempting to understand the strong interactions, physicists were led in the late 1960s and early 1970s to investigate what came to be known as "string theory." String theory was originally discovered by accident, or at least in an exceedingly indirect way, starting with the "Veneziano model" [30]. Surveys of string theory can be found in [17, 11, 25, 13]. As string theory was developed, a remarkably rich mathematical structure emerged, but one which bore increasingly little resemblance to strong interactions. By about 1973-74, a successful theory of strong interactions emerged in the context of nonabelian gauge theory. The mathematical structure of string theory retained its fascination, however. By around 1974, just as the original motivation for work on string theory was fading, it was suggested that string theory should be viewed not as a theory of strong interactions but as a framework for reconciling gravitation with quantum mechanics [24]. This idea has many bizarre implications. For instance, it is necessary to believe that (insofar as the conventional concepts of geometry are valid) space-time is ten-dimensional rather than four-dimensional. After some years of neglect, this idea has been revived in the 1980s, and there are many indications that this framework is close to the truth.

The roundabout path to the discovery of string theory has had a price. Despite learning much about this subject, we still do not know the logical framework in which it has its proper home. It is roughly as if general relativity had been invented, in some peculiar formulation, without knowing about Riemannian geometry; the task would then arise of reconstructing Riemannian geometry as the basic framework behind general relativity. The idea of knowing about general relativity without knowing about Riemannian geometry may sound outlandish, but we are in just such an outlandish situation in string theory. We do not know what the basic logical setting for string theory will turn out to be. We can say that some of the ingredients in string theory are Riemann surfaces, modular forms, and representation theory of infinite-dimensional Lie algebras. These are preliminaries for thinking about string theory just as a modicum of elementary linear algebra is a prerequisite for Riemannian geometry and general relativity. While we do not know the proper logical setting for string theory, it seems rather clear that it will involve some fundamental generalization of the usual concepts

of geometry. This generalization of geometry is bound to have widespread repercussions for mathematics as well as physics. The unearthing of it will entail a new golden age in the interaction of mathematics and physics.

Very probably, in some suitable sense, the number of fundamental mathematical problems is infinite. On the other hand, I personally believe that the number of really fundamental physics problems is finite. If this is so, then there will only be a finite number of episodes in the future in which mathematics and physics will interact in a really fundamental way. It seems likely that the next several decades will be one of those periods.

1. Particle physics in the 1980s. This article is written in four sections. In this section, I will review the basic ingredients in our present knowledge of fundamental physics. In the next section, I will try to explain why the idea that space-time is ten-dimensional (this is one of the requirements of string theory) is not only compatible with everyday experience but even attractive. Surveys of the subjects treated in these two sections can be found in [35, 5, 13]. In the third section, I will sketch a very brief introduction to quantum field theory, emphasizing features that are relevant to string theory. The last section will be devoted to string theory.

We will begin our review of theoretical physics with general relativity. In this theory, space-time is a pseudo-Riemannian manifold M , of signature $(- + \cdots +)$. In this section, M is four-dimensional. I will denote local space-time coordinates as x^i , $i = 1, \dots, 4$. General relativity is governed by a variational principle associated with the Lagrangian

$$S_{\text{GR}} = \frac{1}{16\pi G} \int_M R \quad (1)$$

where R is the Ricci scalar of M and G is Newton's constant. The variational (Euler-Lagrange) equation derived from (1) is the equation

$$R_{ij} = 0 \quad (2)$$

for vanishing of the Ricci tensor R_{ij} . From the fundamental natural constants, G , \hbar (Planck's constant), and c (the speed of light), we can form a quantity with dimensions of mass:

$$M_{\text{Pl}} = \sqrt{\hbar c / G}. \quad (3)$$

This mass, called the Planck mass, is the really natural mass scale in physics. In conventional units, its numerical value is roughly

$$M_{\text{Pl}} \approx 1.02 \times 10^{-5} \text{ grams}. \quad (4)$$

From the fundamental constants we can likewise construct a fundamental length

$$R_{\text{Pl}} = \hbar / M_{\text{Pl}} c \approx 10^{-33} \text{ centimeters}, \quad (5)$$

which is known as the Planck length, and a time

$$t_{\text{Pl}} \approx 10^{-43} \text{ second} \quad (6)$$

called the Planck time. The constants \hbar , c , and G are so fundamental in physics that it is most natural to work in units in which $\hbar = c = G = M_{\text{P1}} = R_{\text{P1}} = t_{\text{P1}} = 1$. In such units, any physical quantity—any length, time, or mass—is simply a number.

Now, the actual values of the fundamental mass, length, and time are very strange. The Planck mass (4) is actually a macroscopic mass (the mass of a bacterium, perhaps), and is totally off the scale of masses of known elementary particles. The electron mass is 10^{22} times smaller than the Planck mass, and the heaviest elementary particles that we are able to produce in accelerators are still 10^{17} times lighter than the Planck mass. (5) and (6) are likewise completely off the usual scale of elementary particle physics. Everything that we know about quantum field theory comes from experiments probing length scales of at least 10^{-16} cm or times of at least 10^{-26} sec. Such lengths and times are very small by ordinary standards, of course, but by an appropriate yardstick determined by the fundamental quantities of physics they are very large. To make direct experimental probes of how nature reconciles quantum mechanics with general relativity would require experiments sensitive to processes that occur on times of order t_{P1} or lengths of order R_{P1} , or with individual elementary particles accelerated to kinetic energies of order M_{P1} . This is regrettably out of reach for the foreseeable future. We can hope for indirect clues from experiment, but progress with quantum gravity will require a great deal of theoretical luck and insight.

Actually, the large value of M_{P1} has consequences visible in everyday life. Saying that M_{P1} is very large compared to the mass m of an ordinary particle is the same as saying that Newton's constant G is very tiny on a scale determined by \hbar , c , and m :

$$G = \hbar c / M_{\text{P1}}^2 \ll \hbar c / m^2. \quad (7)$$

This smallness of Newton's constant means that gravitational forces among individual particles of mass m are very tiny. For interactions among ordinary atoms, gravitation becomes significant only when one considers an aggregation of matter so gigantic that the cumulative effect of gravitational forces among many particles overpowers the extreme weakness of gravity at the atomic level. This means that bodies—such as planets or stars—that form gravitationally out of ordinary atoms must be very large. The fact that the length scale of astronomy is so large compared to the length scale of atoms has the same origin as the fact that the length scale of atoms is so large compared to the Planck length. Both facts are mysteries.

The ordinary masses are so small compared to the natural mass scale of physics that there must be a natural idealization in which they are zero. One of our goals in this section will be to elucidate this generalization.

I would now like to briefly discuss the physical content of general relativity. In two space-time dimensions, the integral (1) is a topological invariant (the Euler characteristic of space-time) and the theory determined by (1) alone does not have much content. In three space-time dimensions, the variational equation

$R_{ij} = 0$ implies that space-time is flat (since on a three-dimensional manifold the whole Riemann tensor can be written in terms of the Ricci tensor). The characteristic features of general relativity first appear in four dimensions. In four dimensions, the Einstein equation $R_{ij} = 0$ does not by any means imply that space-time is flat. On the contrary, this equation has wave-like solutions; if η_{ij} is the flat space Lorentz metric ($\eta = \text{diag} - + \cdots +$) then we can look for a nearly flat solution

$$g_{ij} = \eta_{ij} + h_{ij}, \quad (8)$$

with h considered small. To lowest order in h , the Einstein equations have plane wave solutions

$$h_{ij} = \varepsilon_{ij} e^{ik \cdot x} + \text{complex conjugate}, \quad (9)$$

where k_i and ε_{ij} are constants, obeying

$$k_i k^i = k^i \varepsilon_{ij} = \varepsilon_i^i = 0. \quad (10)$$

The solution (9) is rather analogous to the plane wave solutions of Maxwell's equations, which describe light waves. When the solution (8) of the linearized Einstein equations was discovered, immediately after the formulation of general relativity, this was interpreted as a prediction of the existence of gravitational waves—which should travel at the same speed as light waves because they are both governed by $k_i k^i = 0$. At the time, particles (or “matter”) and waves were interpreted as two very different things, and it was definitely a new kind of wave, not a new kind of particle, that was predicted by general relativity.¹ Ten years later, however, quantum mechanics was developed, and it became clear that waves and particles are different sides of the same coin—the same basic entity will appear as a wave or as a particle depending on the circumstances. Thus general relativity is not only a theory of gravitational forces; it also describes a definite kind of “matter.” On a conceptual plane this is a remarkable triumph. Merely in trying to invent a theory of gravitational forces based on Riemannian geometry, Einstein was forced to invent a unified theory of gravity and matter. A few things are missing, however. General relativity does not seem to make sense as a quantum theory, and the forms of matter observed in nature are richer than what is predicted by general relativity.

Our next task, then, is to discuss some of the other forms of matter (or equivalently, some of the other types of wave) observed in nature. First of all, we have nonabelian gauge forces. Thus, the space-time manifold M , apart from a Riemannian metric, is endowed with additional structure. Over M we have a

¹At the time this prediction was made, and for many decades thereafter, the prospect of actually testing this prediction experimentally seemed hopelessly remote—because of the extreme weakness of the gravitational force. However, the invention of radio astronomy and the discovery of radio pulsars has in the last few years made possible an indirect but compelling experimental test of the theory of gravitational waves.

principle bundle X

$$\begin{array}{c} X \\ \downarrow G \\ M \end{array} \quad (11)$$

with a structure group G that is known by physicists as the “gauge group.” Given any representation R of G , there is an associated vector bundle V_R , which will play an important role in our story later. About G , we know experimentally only that it contains $SU(3) \times SU(2) \times U(1)$ as a subgroup, corresponding to the strong, weak, and electromagnetic interactions, respectively.² Let A be a connection on the bundle V and let F be the corresponding curvature two form. The Yang-Mills action (Lagrangian) is then

$$S_{\text{YM}} = -\frac{1}{4e^2} \int_M |F|^2, \quad (12)$$

where

$$|F|^2 = g^{ii'} g^{jj'} \langle F_{ij} | F_{i'j'} \rangle. \quad (13)$$

Here g_{ij} is the space-time metric, and $\langle | \rangle$ is the Cartan-Killing form on the Lie algebra of G . The constant e in (12) is called the Yang-Mills coupling constant. If the group G is not simple, it is possible to generalize (12), introducing a separate coupling constant for each simple factor in the gauge group.

Of course, the metric g that appears in (13) is supposed to be the same as the one in (1), since all this is happening on a single space-time manifold M . We should properly add the Einstein and Yang-Mills Lagrangians and study the combined theory

$$S = S_{\text{GR}} + S_{\text{YM}}. \quad (14)$$

Upon deriving the Euler-Lagrange variational equations, we will find coupled equations for the Yang-Mills and gravitational fields. This is the proper framework for describing the deflection of light by the sun, and various more exotic processes that unfortunately are undetectably weak.

The next major step is to incorporate what physicists call “fermions,” as opposed to the Yang-Mills and gravitational fields which are “bosons.” To this end, we introduce a Clifford algebra, that is, we introduce the “Dirac matrices” Γ^i , $i = 1, \dots, n$, obeying

$$\Gamma^i \Gamma^j + \Gamma^j \Gamma^i = -2g^{ij}, \quad i, j = 1, \dots, n. \quad (15)$$

For even n , the irreducible representation S of the Clifford algebra has dimension $2^{n/2}$, while for odd n it is $2^{(n-1)/2}$. The Clifford module S automatically furnishes a representation of the Lorentz group $SO(1, n-1)$. The representation

²Experiment tells us more directly about the Lie algebra of G than about G itself. When I say that G contains the subgroup $SU(3) \times SU(2) \times U(1)$, I really mean only that the Lie algebra of G contains that of $SU(3) \times SU(2) \times U(1)$; there is no claim about the global form of G . For the same reason, in later comments I will not be very precise in distinguishing different groups that have the same Lie algebra.

in question is called the spinor representation, the Lorentz generators in this representation being

$$\Sigma^{ij} = -\Sigma^{ji} = \frac{1}{4}[\Gamma^i, \Gamma^j]. \quad (16)$$

For odd n , S is an irreducible representation of the Lorentz group, but for even n —the case that will interest us more— S decomposes as

$$S = S_+ \oplus S_- \quad (17)$$

S_+ and S_- are the eigenspaces of the involution

$$\bar{\Gamma} = i^{(n+2)/4} \Gamma^1 \Gamma^2 \dots \Gamma^n. \quad (18)$$

In four dimensions, the representations S_+ and S_- are complex conjugates of each other.

If the second Stiefel-Whitney class of M vanishes, we can define the “spin bundle” of M , which I will call \hat{S} . It is essentially the vector bundle whose fiber at a point $p \in M$ is the Clifford module specified in (15). A physical field ψ which is a section of \hat{S} is called a “spin one-half fermi field.” Leptons (such as electrons) and quarks (from which protons and neutrons are made) are important examples. The phrase “spin one-half” refers to the fact that the weights of the spinor representation of $\text{SO}(1, N-1)$ are half-integral.

In even dimensions, the spin bundle decomposes as

$$\hat{S} = \hat{S}_+ \oplus \hat{S}_- \quad (19)$$

by analogy with (17). A physical field which is a section of \hat{S}_+ or \hat{S}_- is called a fermi field of positive or negative chirality. Positive and negative chirality fermions are often described as being right-handed or left-handed, respectively; if one shines a beam of positive chirality fermions (particles described mathematically as sections of \hat{S}_+) into a block of matter, it will begin to spin in a right-handed sense.

Various principles of quantum field theory, such as the CPT theorem, require that fermi fields should be real. In four dimensions, since S_+ and S_- are complex conjugates of one another, if we introduce a fermi field ψ which is a section of \hat{S}_+ , we must also introduce a fermi field $\tilde{\psi}$ which is a section of \hat{S}_- . The same would be true, for the same reason, in $4k$ dimensions, for any k . Matters would be very different in $4k+2$ dimensions, since in that case S_+ and S_- are both real (or pseudoreal), but we will reserve that discussion for the next section.

If we do have in a given theory a spin one-half fermi field, what equation should it obey? There is a very natural first-order elliptic operator, the Dirac operator D , which maps sections of \hat{S} to sections of \hat{S} . It is defined as

$$D = \Gamma^i D_i, \quad (20)$$

where D_i is the covariant derivative. For the time being, we consider the Dirac equation for fermions that interact with the gravitational field only, so we consider the covariant derivative constructed from the Levi-Civita connection. In even dimensions, D can be decomposed as

$$D = D_+ \oplus D_-, \quad (21)$$

with

$$D_- : \hat{S}_+ \rightarrow \hat{S}_-, \quad D_+ : \hat{S}_- \rightarrow \hat{S}_+. \quad (22)$$

(In other words, D_+ maps sections of \hat{S}_+ to sections of \hat{S}_- , and vice versa.) The Dirac equation is thus

$$D_- \psi_+ = 0, \quad D_+ \psi_- = 0. \quad (23)$$

This actually is what we would usually call the “massless” Dirac equation. It makes sense to introduce an arbitrary complex constant λ and write

$$D_- \psi_+ + \lambda \psi_- = 0, \quad D_+ \psi_- + \lambda^* \psi_+ = 0. \quad (24)$$

(Here λ^* is simply the complex conjugate of λ , so that the second equation is the complex conjugate of the first.) While (23) describes massless waves which travel at the speed of light just like electromagnetic or gravitational waves, (24) describes massive fermions, in fact fermions of mass λ . Now, as I have explained, there are many important cases (like leptons and quarks) in which the mass term is very tiny, less than 10^{-17} in Planck units. More exactly, the cases in which the mass term is absent to such enormous precision are the only cases which we know about, since our technology does not enable us to discover fermions which have masses of order one, if there are any.

In the precise framework that we have been discussing—fermions that couple to the gravitational field only—there is no natural explanation for why the λ term should be absent. To obtain such an explanation—and describe one of the really central observations in physics—we must reintroduce the nonabelian gauge fields that we have temporarily been suppressing. Letting R be any representation of the gauge group G , there is an associated vector bundle V_R . If \tilde{R} is the dual or complex conjugate representation of R , then $V_{\tilde{R}}$ is the dual bundle to V_R . V_R is canonically isomorphic to $V_{\tilde{R}}$ if R is real or pseudoreal, but not if R is complex.

It makes sense now to consider fermi fields which are sections not of S_+ but of

$$W_+ = \hat{S}_+ \otimes V_R. \quad (25)$$

If we do introduce “right-handed” fermions ψ_+ which are sections of W_+ , then in four dimensions the CPT theorem requires us to also introduce “left-handed” fermions which are sections of the complex conjugate bundle

$$\tilde{W}_- = \hat{S}_- \otimes V_{\tilde{R}}. \quad (26)$$

Now, since V_R and $V_{\tilde{R}}$ are endowed with their Yang-Mills connections, we can write Dirac equations,

$$D_- : W_+ \rightarrow W_-, \quad D_+ : \tilde{W}_- \rightarrow \tilde{W}_+. \quad (27)$$

Here

$$W_- = \hat{S}_- \otimes V_R, \quad \tilde{W}_+ = \hat{S}_+ \otimes V_{\tilde{R}}. \quad (28)$$

Formally, the Dirac equation takes the same form as before,

$$0 = D_- \psi_+ = D_+ \psi_- \quad (29)$$

with now the combined Levi-Civita plus Yang-Mills connection. There is a big difference, though, between (29) and the analogous equation (23) in the absence of Yang-Mills fields. The difference is that if the representation R is complex, then we cannot add a mass term to (29). The equation

$$D_- \psi_+ + \lambda \psi_- \quad (30)$$

only makes sense if the representation R is isomorphic to its dual, since $D_- \psi_+$ is a section of $\hat{S}_- \otimes V_R$, while ψ_+ is a section of $\hat{S}_- \otimes V_{\bar{R}}$. Therefore, if fermions transform in a complex representation of the gauge group G , then we can naturally understand why they are so extremely light compared to the natural scale, the Planck scale.

This is precisely what is observed, and it is believed by most physicists to be the proper explanation for why the observed fermions are so very light. Actually, a more precise statement is necessary, since in fact in nature the observed representation R is far from being irreducible. The decomposition of R into irreducible representations of the gauge group $SU(3) \times SU(2) \times U(1)$ seems to contain at least fifteen pieces:

$$R = \bigoplus_{i=1}^{15} R_i. \quad (31)$$

I will not give an explicit description of the R_i now, since this can be done more economically when we discuss "grand unification" presently. It follows, of course, from (31) that

$$\tilde{R} = \bigoplus_{j=1}^{15} \tilde{R}_j. \quad (32)$$

Now, if there were i and j with $R_i \approx \tilde{R}_j$, then the corresponding fermions ψ_{i+} and ψ_{j-} could have masses, since (30) would make sense. However, the observed fermions all have $R_i \neq \tilde{R}_j$, and this, together with the other facts sketched above, explains why they are so light.

Now, let us shift our point of view slightly. There is no reason at all to believe that the fermions that have been discovered experimentally are all of those which exist. Very probably there are many, perhaps even infinitely many, fermions with masses "of order one," that is, of order M_{P1} . Such fermions, of course, must transform in real representations of G . And conversely, fermions in real representations will very plausibly have masses of order one. Let U and \tilde{U} be the G representations of fundamental right- and left-handed fermions in some underlying theory of nature. Of course, they are duals of one another. In the spirit of K theory, form the formal difference of representations:

$$\Delta = U \ominus \tilde{U}. \quad (33)$$

In addition to observed right-handed fermions which transform as R , U may contain additional right-handed fermions transforming in some representation U_0

$$U = U_0 \oplus R \quad (34)$$

which must be real since the U_0 particles have gotten mass. (34) implies

$$\tilde{U} = U_0 \oplus \tilde{R}. \quad (35)$$

Therefore, when we form the representation difference (33), U_0 will cancel out:

$$\Delta = U \ominus \tilde{U} = R \ominus \tilde{R}. \quad (36)$$

We conclude, then, that $R \ominus \tilde{R}$, which we can determine at accelerators, is the same as the underlying representation difference (33), which is a property of the fundamental theory. This is why the quantum numbers of the low energy fermions are so important.

A few questions arise here. First of all, one might feel that we have done too good a job of explaining why the light fermions are light. The quark and lepton masses are not zero; they range from about 10^{-22} to perhaps 10^{-17} in the natural units I described earlier. The framework we have sketched might appear to force the quark and lepton masses to be strictly zero, since a mass term in (30) is strictly impossible. The answer to this involves something called “symmetry breaking.” If we are presented with a vector bundle V with structure group G , it might happen that under some conditions the structure group of V can be reduced to a subgroup G_0 . This phenomenon has an analogue in particle physics; it is called gauge symmetry breaking and plays a central role in the Weinberg-Salam-Glashow mode of weak interactions. For the moment, I will simply assume that to a mathematical audience it is plausible that there can be a “physical” counterpart of reducing the structure group of a vector bundle to a subgroup.

Suppose that at some low mass scale m , the gauge group G is effectively reduced to a subgroup G_0 . Even if the representations R and \tilde{R} are inequivalent as representations of G , they may be equivalent as representations of G_0 . In this case, the fermions that were kept massless by the inequivalence of R and \tilde{R} will be able to gain masses of order m . This is precisely what seems to happen in nature. At a mass scale of order $10^{-17} M_{\text{Pl}}$, the gauge group $\text{SU}(3) \times \text{SU}(2) \times \text{U}(1)$ is reduced to $\text{SU}(3) \times \text{U}(1)$.³ At this point, some of the gauge fields become massive (these being the W and Z particles which were discovered at CERN several years ago—the heaviest elementary particles that have been discovered). At the same time, the representations R and \tilde{R} are isomorphic as representations of $\text{SU}(3) \times \text{U}(1)$, so the light fermions can and do gain mass. In all of this, we do not understand why the mass scale associated with symmetry breaking is so tiny compared to the natural mass scale M_{Pl} . It is, however, pretty clear from our discussion that the idealization in which the masses of the particles are all zero is the situation in which the gauge group $\text{SU}(3) \times \text{SU}(2) \times \text{U}(1)$ is not broken to a subgroup.

In this brief survey of contemporary physics, the idea that the representation R might be complex probably did not seem very shocking. Yet historically,

³These are strong and electromagnetic interactions, respectively. The $\text{U}(1)$ in question is a sort of skewed embedding in the original $\text{SU}(2) \times \text{U}(1)$.

when this was discovered in the 1950s (or more precisely when certain facts were discovered which nowadays have the interpretation I have indicated) it came as a shocking surprise. Why is this? In the basic decomposition $S = S_+ \oplus S_-$ of the spinor representation S into spinors S_{\pm} of positive and negative chirality, the distinction between S_+ and S_- is a matter of convention. Under a change of the orientation of space-time, called a parity transformation by physicists, S_+ and S_- are exchanged. The representations R and \tilde{R} are therefore exchanged by parity. If we assume that the laws of nature are invariant under parity, then R and \tilde{R} must be isomorphic. Our explanation of the lightness of the fermions therefore rests on parity violation. It was considered startling when in the 1950s it was discovered that the weak interactions violate parity. On the other hand, parity is conserved by strong and electromagnetic interactions; this is the statement that R and \tilde{R} are isomorphic as representations of $SU(3) \times U(1)$.

Once we realize that symmetry breaking plays an important role in explaining why the light fermions are not strictly massless, it is natural to carry this thought a step further. We know that the gauge group of nature contains at least $G = SU(3) \times SU(2) \times U(1)$. We also know that at very low energies this is reduced to a subgroup $G_0 = SU(3) \times U(1)$. Looking up to higher energies, might the gauge group G which we observe at accessible energies be itself a reduction of a larger group \bar{G} which is relevant at higher energies? It would be far more satisfying to describe nature not by the ungainly product $SU(3) \times SU(2) \times U(1)$ but by a more simple structure. This is the goal of so-called “grand unified theories.” The most obvious example of a simple group that contains $SU(3) \times SU(2) \times U(1)$ is $SU(5)$. The embedding of $SU(3) \times SU(2)$ in 5×5 matrices of $SU(5)$ can be indicated as follows:

$$\begin{pmatrix} SU(3) & 0 \\ 0 & SU(2) \end{pmatrix}. \quad (37)$$

We then take the $U(1)$ generator to be the unique traceless 5×5 matrix which commutes with the above embedding of $SU(3) \times SU(2)$, namely,

$$\text{diag}(1, 1, 1 - 3/2, -3/2). \quad (38)$$

If we are to embed the observed gauge group $SU(3) \times SU(2) \times U(1)$ in $SU(5)$, then the representation difference $\Delta = R \ominus \tilde{R}$ must have a natural interpretation as a difference of $SU(5)$ representations. Indeed it does, and here I will write down for the first time an explicit formula for the representation difference observed in nature; this would have been rather clumsy at the $SU(3) \times SU(2) \times U(1)$ level. Let $\mathbf{5}$ be the fundamental five-dimensional representation of $SU(5)$, and let $\bar{\mathbf{5}}$ be its dual. Let $\mathbf{10}$ be the antisymmetric part of the tensor product $\mathbf{5} \otimes \mathbf{5}$, and let $\bar{\mathbf{10}}$ be its dual. Then the fermions that we observe in nature come in three copies of a basic structure usually called a “generation.” The left-handed fermions of a given generation are observed to transform as $\bar{\mathbf{5}} \oplus \mathbf{10}$, while the right-handed fermions transform as the dual of this or $\mathbf{5} \oplus \bar{\mathbf{10}}$. In nature we observe three copies of this, so in other words the representation difference which we observe

is

$$\Delta = 3(\bar{\mathbf{5}} \oplus \mathbf{10} \ominus \mathbf{5} \ominus \overline{\mathbf{10}}). \quad (39)$$

When one contemplates extending the observed gauge group G to a larger group \overline{G} , it is important to understand that this would entail predicting new forces. In the case of $SU(5)$, and most other grand unified theories, the new forces have a dramatic consequence: they cause the proton (which is otherwise absolutely stable), to be unstable, with a lifetime typically in the range of 10^{32} – 10^{45} years, depending on which grand unified theory one considers. In 10^4 tons of water, roughly the largest quantity in a feasible experiment, there are about 10^{34} protons, so assuming that an event rate of one proton decay per year is detectable, the lower part of the interesting range can be probed experimentally. So far proton decay has not been observed.

(39) is much nicer than the corresponding formula at the $SU(3) \times SU(2) \times U(1)$ level, and this is one of the main reasons to believe that the $SU(5)$ model probably has some truth in it. However, it is natural to wonder whether there are groups beyond $SU(5)$ that would be as good or better. One group that works very nicely is $SO(10)$, with the unique nontrivial embedding of $SU(5)$ in $SO(10)$. (The fundamental vector of $SO(10)$, which we will call $\mathbf{10}$, decomposes under $SU(5)$ as $\mathbf{5} \oplus \bar{\mathbf{5}}$.) $SO(10)$ has two complex conjugate spinor representations of 16 dimensions each; let us call them $\mathbf{16}$ and $\overline{\mathbf{16}}$. Under $SU(5)$ the $\mathbf{16}$ decomposes as $\mathbf{1} \oplus \bar{\mathbf{5}} \oplus \mathbf{10}$, where $\mathbf{1}$ is the trivial representation of $SU(5)$. This is, of course, a real representation. The $\overline{\mathbf{16}}$ decomposes, of course, as the complex conjugate of the $\mathbf{16}$, or $\mathbf{1} \oplus \mathbf{5} \oplus \overline{\mathbf{10}}$. Therefore, at the $SO(10)$ level, we can rewrite (39) as

$$\Delta = 3(\mathbf{16} \ominus \overline{\mathbf{16}}). \quad (40)$$

This is about as simple as a nontrivial representation difference can be, except that the factor of three is rather peculiar; we will discuss some attempts to explain it in the next section. It is natural, though, to ask whether there are any examples of groups beyond $SO(10)$ that work just as nicely. It turns out that there is one, namely, the exceptional group E_6 . The smallest nontrivial representation of E_6 is a complex representation which we will call the $\mathbf{27}$, this being its dimension; its dual will be denoted as the $\overline{\mathbf{27}}$. E_6 has an $SO(10)$ subgroup, the decomposition of the $\mathbf{27}$ under $SO(10)$ being $\mathbf{27} = \mathbf{16} \oplus \mathbf{10} \oplus \mathbf{1}$, with $\mathbf{16}$ and $\mathbf{10}$ being as before the spinor and vector of $SO(10)$. At the E_6 level, we can then write

$$\Delta = 3(\mathbf{27} \ominus \overline{\mathbf{27}}). \quad (41)$$

This is the third and last example of a group that seems suitable for grand unification in four dimensions. Having found a hint that E_6 can play a role, it is natural to wonder whether one can go further and base a grand unified theory on the biggest exceptional group E_8 . This runs into the problem that E_8 only has real representations, and so would automatically lead to $\Delta = 0$. It is possible to try to use E_8 for grand unification, but this requires some additional ingredients, which we will discuss in the next section.

If one wants to summarize our knowledge of physics in the briefest possible terms, there are three really fundamental observations: (i) Space-time is a pseudo-Riemannian manifold M , endowed with a metric tensor and governed by geometrical laws. (ii) Over M is a vector bundle X with a nonabelian gauge group G . (iii) Fermions are sections of $(\hat{S}_+ \otimes V_R) \oplus (\hat{S}_- \otimes V_{\tilde{R}})$. R and \tilde{R} are not isomorphic; their failure to be isomorphic explains why the light fermions are light and presumably has its origins in a representation difference Δ in some underlying theory. All of this must be supplemented with the understanding that the geometrical laws obeyed by the metric tensor, the gauge fields, and the fermions are to be interpreted in quantum mechanical terms.

2. Physics in ten dimensions. The standard model of physics which we have just surveyed has many successes, but leaves many questions open. Surely we would like to find a natural explanation for the peculiar factor of three in (40) and (41). We do not want to merely assume that nature is based on three copies of $\mathbf{16} \oplus \overline{\mathbf{16}}$ of $\text{SO}(10)$ or three copies of $\mathbf{27} \oplus \overline{\mathbf{27}}$ of E_6 . There must be some more economical structure at a more elementary level. Also, it is very peculiar to study a chain that passes from $\text{SU}(5)$ to $\text{SO}(10)$ to E_6 without finding some way to extend this to E_8 . We would like some natural understanding of the symmetry-breaking steps along this chain. Finally, the fact that all we observe in experiment is the character *difference* (33) or (36) is a warning that the true underlying structure may be far richer than is apparent. The underlying representation U may even be infinite-dimensional; of course, in this case there will have to be some suitable regularization in the definition of the character difference $\Delta = U \ominus \tilde{U}$.

I will now describe an approach to some of these questions. Let us assume that space-time is not a *four*-dimensional pseudo-Riemannian manifold, but has a higher dimension; in fact, the case of ten dimensions is favored in string theory. We consider thus a *ten*-dimensional manifold \overline{M} , oriented, with a spin structure, and with a metric of signature $(- + + + \cdots +)$. One ingredient in the ten-dimensional theory will be the Einstein action,

$$S_{\text{GR}} = \frac{1}{16\pi G} \int_{\overline{M}} R. \quad (42)$$

At first sight, one might believe that it is preposterous to imagine that the world is ten-dimensional. The world seems to be “obviously” four-dimensional (if one includes time along with the three obvious space dimensions). The following is crucial, however. We tend to take for granted that there is some notion of “empty space” and that any other physical state is obtained by adding particles to “empty space.” In reality, though, what we interpret as empty space is just some solution of the underlying physical equations which plays a special role in our experience. The reason that there is a solution playing such a special role is largely that the energy scales of our experiments are so small compared to the Planck mass M_{Pl} . The tools at our disposal are so feeble that we can bring about only minor disturbances in whatever solution of the underlying equations

we happen to have been born into. Just to make this discussion more concrete, suppose that the ten-dimensional Einstein equations derived from (42) are the fundamental equations of physics. Let M^4 be flat four-dimensional Minkowski space, and let K be some compact Ricci-flat six manifold. Suppose that the “vacuum state” of the ten-dimensional world is

$$\overline{M} = M^4 \times K. \quad (43)$$

Suppose finally that the radius of K (by which I mean any characteristic measure of the size of K) is comparable to the typical length scale of physics, namely R_{Pl} . This is so tiny compared to the length scales of our observations that K will be indistinguishable from a point; the existence of extra dimensions will not be obvious in everyday life, or even in accelerator experiments. Thus, \overline{M} will be indistinguishable from four-dimensional Minkowski space; likewise, we could imitate a four-dimensional cosmological model.

This may seem reasonable intuitively. To make things more precise, we will now discuss an important aspect of ten-dimensional physics, namely, the ten-dimensional Dirac equation. We thus introduce the ten-dimensional Dirac operator

$$D_{(10)} = \sum_{i=1}^{10} \Gamma^i D_i. \quad (44)$$

We can write it as

$$D_{(10)} = D_{(4)} + D_K, \quad (45)$$

where $D_{(4)}$ is the four-dimensional Dirac operator

$$D_{(4)} = \sum_{i=1}^4 \Gamma^i D_i \quad (46)$$

and $D_{(K)}$ is the Dirac operator of K ,

$$D_K = \sum_{j=5}^{10} \Gamma^j D_j. \quad (47)$$

We would now like to analyze the ten-dimensional Dirac equation

$$0 = D_{(10)} \Psi(x^i, y^j), \quad (48)$$

where x^i and y^j , $i = 1, \dots, 4$, $j = 5, \dots, 10$, are coordinates of M^4 , and Ψ is a spinor field in ten dimensions.

Let us first briefly discuss just what kind of spinor Ψ will be. The irreducible representation of the Clifford algebra in ten dimensions decomposes as the sum of two irreducible representations of the Lorentz group $\text{SO}(1, 9)$. They are distinguished by the eigenvalue of

$$\Gamma^{(10)} = \Gamma^1 \Gamma^2 \dots \Gamma^{10}. \quad (49)$$

In ten dimensions, unlike four dimensions, the two irreducible spin representations of $\text{SO}(1, 9)$, which I will call $S_+^{(10)}$ and $S_-^{(10)}$, are both real. So the CPT

theorem permits us to consider a theory in which the spinor field Ψ transforms according to one definite spin representation, say $S_+^{(10)}$. In other words

$$\Gamma^{(10)}\Psi = +\Psi. \quad (50)$$

Actually, ten-dimensional supersymmetry and string theory force us to consider this case. If we are interested in thinking about four-dimensional physics, we should decompose the spinor representation of $SO(1,9)$ under $SO(1,3) \times SO(6)$, where $SO(1,3)$ is the Lorentz group of M^4 and $SO(6)$ is the structure group of the tangent bundle of K . In this decomposition, only spinor representations of $SO(1,3)$ and $SO(6)$ will appear, since the representation space for a ten-dimensional Clifford algebra certainly represents the four-dimensional and six-dimensional Clifford subalgebras. The precise decomposition is easily worked out by introducing the operators analogous to (49):

$$\Gamma^{(4)} = i\Gamma^1\Gamma^2 \dots \Gamma^4, \quad \Gamma^{(K)} = -i\Gamma^5\Gamma^6 \dots \Gamma^{10}. \quad (51)$$

(The factors of $\pm i$ are conventional and ensure that the eigenvalues of $\Gamma^{(4)}$ and $\Gamma^{(K)}$ are ± 1 .) These matrices obey the obvious relation

$$\Gamma^{(10)} = \Gamma^{(4)} \cdot \Gamma^{(K)}. \quad (52)$$

Therefore, in an irreducible representation of $SO(1,9)$ which has $\Gamma^{(10)} = +1$, $\Gamma^{(4)}$ and $\Gamma^{(K)}$ are equal,

$$\Gamma^{(4)} = \Gamma^{(K)}. \quad (53)$$

This means that the decomposition of the $SO(1,9)$ spinor representation $S_+^{(10)}$ of positive 'chirality' under $SO(1,3) \times SO(6)$ is

$$S_+^{(10)} = \left(S_+^{(4)} \otimes S_+^{(K)}\right) \oplus \left(S_-^{(4)} \otimes S_-^{(K)}\right). \quad (54)$$

Here $S_\pm^{(4)}$ and $S_\pm^{(K)}$ are the positive and negative chirality spin representations of $SO(1,3)$ and $SO(6)$, respectively.

Going back to our problem, we want to solve (48) by separation of variables, using the decomposition (45). Since $D_{(4)}$ and D_K do not commute, but anticommute, the standard procedure of separation of variables needs slight modification.

It is convenient to introduce

$$D'_K = \Gamma^{(4)} D_K \quad (55)$$

which does commute with $D_{(4)}$. D'_K is unitarily equivalent to D_K , and so has the same spectrum (since the matrices $\Gamma^{(4)}\Gamma^j$, $j = 5, \dots, 10$, obey the same Clifford algebra as Γ^j , and the irreducible representation of this algebra is known to be unique). Introduce a complete set of eigenfunctions χ_m of D'_K ,

$$D'_K \chi_m = \lambda_m \chi_m. \quad (56)$$

We then write

$$\Psi(x^i, y^j) = \sum_m \phi_m(x^i) \otimes \chi_m(y^j) \quad (57)$$

whereupon (48) reduces to

$$0 = (D_{(4)} + \Gamma^{(4)}\lambda_m)\psi_m \quad (58)$$

or equivalently

$$0 = (D'_{(4)} + \lambda_m)\psi_m, \quad (59)$$

where we have introduced

$$D'_{(4)} = \Gamma^{(4)}D_{(4)}. \quad (60)$$

$D'_{(4)}$ is unitarily equivalent to $D_{(4)}$ (since $\Gamma^{(4)}\Gamma^i$ generate a Clifford algebra). (59) is equivalent to the Dirac equation for a massive fermion introduced in (24).

We have thus learned the following important lesson. The eigenvalues λ_m of the Dirac operator D_K on the compact manifold K correspond in four-dimensional terms to the masses of the fermions ψ_m . Of course, there are an infinite number of such eigenvalues. But as D_K is an elliptic operator on a compact manifold, there are only a finite number of zero eigenvalues—the eigenvalue zero only appears with a finite multiplicity. The nonzero eigenvalues of D_K will be of order $1/R$, with R being the radius of K . Since we are assuming that the radius of K is of order the Planck length (5), the nonzero eigenvalues of D_K will correspond to fermions with Planckian masses—fermions that we would certainly not be able to discover experimentally. Experimentally accessible four-dimensional physics will be determined by the zero eigenvalues of D_K , corresponding to massless particles in four dimensions. As there are only finitely many of these, the ten-dimensional theory will look for all practical purposes like a four-dimensional theory with a finite number of fermi fields. This is just the sort of structure that we discussed in the last section, so we have achieved our goal of showing that a theory that is really ten-dimensional can look four-dimensional to an observer whose experiments are limited to low energies.

What is more, it is very interesting that in this framework the observed fermions in four dimensions originate as zero modes of the Dirac operator D_K . Zero eigenvalues of elliptic operators such as the Dirac operator do not arise by accident; they arise when they are related to suitable topological invariants. So here (and in many other instances) basic physical questions lead to questions about the topology of K .

The simplest topological invariant that can predict the existence of zero modes of an elliptic operator is the *index* of the operator. Let n_+ and n_- be the number of zero eigenvalues of D_K of positive and negative chirality, respectively (i.e., $\Gamma_K = \pm 1$). Then the difference $n_+ - n_-$ is called the *index* of the Dirac operator, and is easily shown to be a topological invariant. We have so far been tacitly considering a ten-dimensional Dirac equation for a spinor field coupled to the space-time geometry only—no Yang-Mills vector bundle. In this case, it is easily seen from the Atiyah-Singer index theorem, or on various more elementary grounds, that in six dimensions (and more generally in $4k + 2$ dimensions) the index of the Dirac operator vanishes. This should come as no surprise; in the last section we found a rationale for the existence of massless fermions in four

dimensions only in the case in which there is a Yang-Mills group. Let us therefore generalize our discussion and reintroduce the Yang-Mills gauge fields.

The preceding considerations are rather general. The case favored by developments of the last few years [12, 14] in string theory is a ten-dimensional theory with gauge group $E_8 \times E_8$. For brevity I will consider only a single E_8 . The fermions will be in the adjoint representation of E_8 . In a ten-dimensional theory which has Yang-Mills fields as well as the gravitational field, to describe the vacuum state it is not enough to specify the vacuum manifold. It is also necessary to specify an E_8 vector bundle X over space-time, endowed with a connection A .

What would be a natural choice of X ?⁴ Whenever we discuss a six manifold K , there is one vector bundle that is always present—the tangent bundle T . This of course is endowed with the Levi-Civita connections. The structure group of T is—in the general case— $SO(6)$. Any embedding of $SO(6)$ in E_8 gives a canonical way to construct an E_8 bundle with connection from the tangent bundle T with its Riemannian connection. There is one embedding of $SO(6)$ in E_8 which is in a sense minimal among such embeddings. It comes from the chain

$$SO(6) \times SO(10) \subset SO(16) \subset E_8. \quad (61)$$

Here $SO(16)$ is a maximal subgroup of E_8 , and $SO(6) \times SO(10)$ is a maximal subgroup of $SO(16)$.⁵ The embedding (61) turns out to lead to an interesting picture of four-dimensional physics.

At this point we encounter the notion of gauge symmetry breaking, alluded to in the last section. What will a low energy physicist interpret as the gauge group? A low energy physicist is “trapped” in a world with an E_8 bundle X and some particular connection A , and is not able to disturb this world very much. An E_8 gauge transformation that does not leave A invariant is not a symmetry of this particular world but relates it to another world with some other connection A' . Probing such a gauge transformation would involve exciting the very massive degrees of freedom whose inaccessibility to the low-energy observer is the reason that the world is apparently four-dimensional. What the low-energy physicist interprets as the gauge group is the subgroup of E_8 that acts on the bundle X , leaving the connection A invariant. For a generic choice of X and A , this group would be trivial. If, however, X is constructed from the tangent bundle via the embedding (61) of $SO(6)$ in E_8 , then the subgroup of gauge transformations that leaves the connection invariant is the group $G = SO(10)$ that commutes with $SO(6)$.

This is a promising development, because although E_8 is not a suitable gauge group for a four-dimensional theory (it only has real representations and so would lead to $\Delta = 0$), $SO(10)$ is one of the natural candidates, as we learned in the last section. Let us now compute the character difference Δ which will emerge

⁴The following construction was considered in [31, 4].

⁵Recall that we are really working at the Lie algebra level, and not specifying the global structure of the various groups. $SO(16)$ in (61) is really $\text{spin}(16)/\mathbb{Z}_2$, etc.

in the present framework. To do so, it is necessary to decompose the adjoint representation of E_8 under $SO(6) \times SO(10)$. The adjoint representation of E_8 , which we will call the **248**, decomposes under $SO(6) \times SO(10)$ in the general form

$$\mathbf{248} = \sum_i L_i \otimes R_i, \quad (62)$$

where L_i and R_i are certain representations of $SO(6)$ and $SO(10)$, respectively. For each L_i , there is a corresponding Dirac operator $D_K^{(i)} = D_K^{L_i}$ acting on fermions that transform as L_i under $SO(6)$. Massless fermions in four dimensions that transform as R_i under $SO(10)$ originate as zero modes of $D_K^{(i)}$. In view of (53), zero eigenvalues of $D_K^{(i)}$ with $\Gamma^K = +1$ (or -1) have $\Gamma^{(4)} = +1$ (or -1). Let δ_i be the index of $D_K^{(i)}$. It is the difference between the number of zero eigenvalues of $D_K^{(i)}$ with $\Gamma^K = \pm 1$ or equivalently the difference between the number of zero eigenvalues of $D_K^{(i)}$ with $\Gamma^{(4)} = \pm 1$. Therefore, in the basic character difference Δ of the massless fermions, the $SO(10)$ representation R_i will appear with the coefficient δ_i . Altogether, the character difference Δ will be

$$\Delta = \sum_i \delta_i R_i. \quad (63)$$

We will now evaluate (63) in detail. Let us adopt some conventions. If M is an $SO(6)$ representation and N is an $SO(10)$ representation, the tensor product $M \otimes N$ will be denoted (M, N) . Representations of $SO(6)$ and $SO(10)$ will be labeled by their dimension. The relevant representations of $SO(6)$ are the adjoint, **15**, the vector, **6**, and the two spinor representations, **4** and $\bar{\mathbf{4}}$. The relevant representations of $SO(10)$ are the adjoint, **45**, the vector **10**, and the two spinor representations **16** and $\bar{\mathbf{16}}$. The adjoint representation of E_8 decomposes under $SO(6) \times SO(10)$ as

$$\mathbf{248} = (\mathbf{15}, \mathbf{1}) \oplus (\mathbf{1}, \mathbf{45}) \oplus (\mathbf{6}, \mathbf{10}) \oplus (\mathbf{4}, \mathbf{16}) \oplus (\bar{\mathbf{4}}, \bar{\mathbf{16}}). \quad (64)$$

If L_i is a real representation of $SO(6)$, then (from the Atiyah-Singer index theorem or various more elementary considerations) the index δ_i is zero. What is more, if L_i and $L_{\bar{i}}$ are complex conjugate representations, then $\delta_i = -\delta_{\bar{i}}$. The only complex representations of $SO(6)$ in (64) are the **4** and $\bar{\mathbf{4}}$. So (63) reduces to

$$\Delta = \delta_4 (\mathbf{16} - \bar{\mathbf{16}}). \quad (65)$$

Comparing to (40), we see that this is of the correct form to agree with observations, with δ_4 being ± 3 in nature as far as we can see.⁶

As for the actual value of δ_4 , it is one of the most fundamental topological invariants of a six manifold. The de Rham complex of differential forms can be built from the tensor product of two spin bundles. Since the **4** of $SO(6)$ is one of the two spinor representations, a spinor on K with values in the **4** of $SO(6)$ is equivalent to a certain collection of differential forms. δ_4 can therefore be

⁶We cannot determine the sign, since the difference between **16** and $\bar{\mathbf{16}}$ of $SO(10)$ is a matter of convention.

related to the Euler characteristic χ and the Hirzebruch signature σ , which are the topological invariants that can be made from an index problem in the de Rham complex. Actually, $\sigma = 0$ in six (or $4k + 2$) dimensions, and

$$\delta_4 = \chi/2. \quad (66)$$

Thus, we have seen how the observed character difference Δ can be related to something more fundamental, namely, the topology of K . Although we have not succeeded in explaining the peculiar number 3 in (40), we have perhaps removed some of the mystery from this. The reason that nature seems to repeat herself, with several “fermion generations,” is simply that there is no reason for a six-dimensional manifold to have Euler characteristic ± 2 . A suitable K , starting as we have done with a single spinor field Ψ in ten dimensions, can give rise to any desired number of fermion generations in four dimensions.

I have tried to give the flavor of how properties of four-dimensional physics can be extracted from geometric and topological properties of K . There are various other examples, but these should suffice as illustrations. I would like to emphasize, though, that while we have supposed the *vacuum* to be a product $M^4 \times K$, the general physical disturbances do not preserve this product structure. The basic laws are supposed to be ten-dimensional laws, governed by (42) or (more likely) some much more refined ten-dimensional theory, and an approximate four-dimensional picture arises only because the “vacuum” admits four-dimensional but not ten-dimensional Poincaré symmetry. Why this is, and what K should be, remain mysteries.

3. Quantum field theory on a Riemann surface. In the last two sections, we have sketched some of the key ingredients in physics in *classical* terms. Buried in the fine print was the proviso that Yang-Mills theory, etc., should actually be reinterpreted quantum mechanically. This in fact leads to a vastly richer and more formidable structure.⁷

Quantum field theory has not yet emerged as an important tool in pure mathematics. But there are indications that this will change in the coming period. Both in the theory of affine Lie algebras and in algebraic geometry, structures that are familiar in quantum field theory have recently come to play a major role. (We will hear about the latter subject from G. Faltings in his lecture at the 1986 International Congress of Mathematicians, “Neuere Entwicklungen in der arithmetischen algebraischen Geometrie.”) In trying here to give a very brief introduction to quantum field theory, I will emphasize those aspects of this subject which are necessary for understanding string theory, and which are likely to be related to the areas of mathematics which I have just mentioned.

Let Σ be a Riemann surface, perhaps with boundary, and let ϕ be a real-valued function on Σ , that is, a map from Σ to R (the real numbers). Let

$$I(\phi) = \frac{1}{2} \int_{\Sigma} d\phi \wedge *d\phi. \quad (67)$$

⁷See, e.g., [16, 10] for introductions.

This is called “the action functional of free boson field theory.” Let f be a real-valued function on $\partial\Sigma$ (the boundary of Σ), and let $\Omega_f(\Sigma)$ be the space of continuous real-valued functions

$$\phi: \Sigma \rightarrow R \quad (68)$$

whose restriction to $\partial\Sigma$ coincides with f .

If we are given an actual metric on Σ (and not just a conformal class of metrics), then the affine space $\Omega_f(\Sigma)$ acquires a natural Riemannian metric. For $\phi, \phi' \in \Omega_f(\Sigma)$, one defines

$$|\phi - \phi'|^2 = \int_{\Sigma} (\phi - \phi')^2. \quad (69)$$

In finite dimensions, a Riemannian metric always induces a measure (“the square root of the determinant of the metric”). Hoping that this still works in infinite dimensions, we can try to define

$$Z_f(\Sigma) = \int_{\Omega_f(\Sigma)} e^{-I(\phi)}. \quad (70)$$

Let us see what is involved in trying to define (70). $I(\phi)$ is a quadratic functional of ϕ , so the integral that we are trying to do is rather similar to a finite-dimensional Gaussian integral

$$Z(A) = \int_{-\infty}^{\infty} d\phi_1 d\phi_2 \cdots d\phi_n e^{-(\phi, A\phi)/2}. \quad (71)$$

Here ϕ_i are coordinates in an n -dimensional Euclidean space, and A is a positive definite quadratic form in n variables. It is well known that the value of (71) is

$$Z(A) = (2\pi)^{n/2} \frac{1}{\sqrt{\det A}} = (2\pi)^{n/2} \prod_i \lambda_i^{-1/2}, \quad (72)$$

with λ_i being the eigenvalues of A . Trying to generalize (72) to infinite dimensions, the factor of $(2\pi)^{n/2}$ does not look very promising for $n \rightarrow \infty$. For this reason, and because of the difficulty in defining the determinant $\det A$ in infinite dimensions, we will not be able to define the integral $Z_f(\Sigma)$ for given f and Σ , but ratios such as $Z_{f_1}(\Sigma_1)/Z_{f_2}(\Sigma_2)$ will make sense.

What is the analogue of A in (70)? Evidently, $I(\phi)$ as defined in (67) is equivalent to

$$I(\phi) = \frac{1}{2} \int_{\Sigma} \phi \Delta \phi, \quad (73)$$

where $\Delta = *d*d$ is the usual Laplacian. Let us consider first the case in which Σ has no boundary. Then Δ is a positive definite operator with a discrete spectrum

$$\Delta \phi_i = \lambda_i \phi_i, \quad i = 1, 2, 3, \dots \quad (74)$$

There is one zero eigenvalue, corresponding to the constant function 1, and the others are positive. The integral in (70) should then be precisely an infinite-dimensional analogue of (71), so we have to define a regularized version of

$$\prod_i \lambda_i^{-1/2}. \quad (75)$$

Before trying to define this infinite product, it is necessary to remove the zero eigenvalue. We try to define

$$\prod_{\lambda_i \neq 0} \lambda_i^{-1/2}. \quad (76)$$

This may be done with zeta function regularization. Let

$$\zeta(s) = \sum_{\lambda_i \neq 0} \lambda_i^{-s}. \quad (77)$$

The series converges if the real part of s is large enough. It defines a meromorphic function of s which can be shown to be regular at $s = 0$. Then we define

$$\prod_{\lambda_i \neq 0} \lambda_i^{-1} = \exp(\zeta'(0)/2). \quad (78)$$

For $\partial\Sigma = 0$, we define then

$$Z(\Sigma) = \exp(\zeta'(0)/2). \quad (79)$$

Now, let us work out a formula analogous to (79) for the more general case $\partial\Sigma \neq 0$. The first case to consider is $f = 0$. This hardly presents any novelty. If $f = 0$, then (70) involves an integral over functions ϕ that vanish on $\partial\Sigma$. With the boundary condition $\phi|_{\partial\Sigma} = 0$, the Laplacian Δ has a discrete spectrum of positive eigenvalues, so we define as before

$$Z_{f=0}(\Sigma) = \exp(\zeta'(0)/2). \quad (80)$$

With $\partial\Sigma \neq 0$, and boundary conditions that $\phi = 0$ on $\partial\Sigma$, the operator Δ is strictly positive definite, so there is no need to remove zero eigenvalues in defining (80). It remains to generalize (80) to $f \neq 0$. To accomplish this, we recall the classical theorem that the functional $I(\phi)$ has a unique extremum subject to the boundary condition $\phi|_{\partial\Sigma} = f$. Denote the extremizing function as ϕ_0 , and write $\phi = \phi_0 + \phi'$. Since $I(\phi)$ is a quadratic functional of ϕ , and stationary at $\phi = \phi_0$, we have

$$I(\phi) = I(\phi_0) + I(\phi'). \quad (81)$$

If ϕ and ϕ_0 are in $\Omega_f(\Sigma)$, then $\phi' = \phi - \phi_0$ is in $\Omega_0(\Sigma)$. So looking back to (70), the change of variables from ϕ to ϕ' converts an integral over $\Omega_f(\Sigma)$ into an integral over $\Omega_0(\Sigma)$, giving

$$Z_f(\Sigma) = e^{-I(\phi_0)} \cdot Z_0(\Sigma) = e^{-I(\phi_0)} e^{\zeta'(0)/2}. \quad (82)$$

Of course, as we have formulated things, (79) and (82) are just definitions. They become theorems only in the context of a suitable infinite-dimensional integration theory.

Now, a variety of comments are in order here. First of all, in the above we have been discussing *real-valued* functions on Σ —that is, maps $\Sigma \rightarrow \mathbb{R}$. More generally, we could pick a Riemannian manifold X , with metric tensor γ , and consider maps from Σ to X . Fixing a map $f: \partial\Sigma \rightarrow X$, let $\Omega_f(\Sigma; X)$ be the space of continuous maps from Σ to X whose restriction to $\partial\Sigma$ coincides with f .

Picking local coordinates ϕ_i on X , a map $\Phi: \Sigma \rightarrow X$ can be described locally in terms of real-valued functions ϕ^i on Σ . We generalize (67) to

$$I(\Phi) = \int_{\Sigma} \gamma_{ij}(\Phi) d\phi^i \wedge *d\phi^j. \quad (83)$$

Instead of (70), we want to define

$$Z_f(\Sigma; X) = \int_{\Omega_f(\Sigma; X)} e^{-I}. \quad (84)$$

Now, (70) was an integral over an affine space, and could be defined explicitly as in (79) and (82). (70) is one way of defining what physicists call (bosonic) free field theory. If X is not flat, (84) is an integral over a nonlinear space, and is rather difficult to deal with. (84) is an example of what physicists would call an interacting or nonlinear quantum field theory. The quantum field theory defined in (84) is known as the nonlinear sigma model, and is particularly important in string theory. To define (84) requires much more than the ζ function regularization that we used in free field theory. In fact, (84) cannot be satisfactorily defined for arbitrary X . Generally speaking, (84) can be defined for manifolds X of positive or zero Ricci tensor but not for manifolds of negative Ricci tensor. Much is known about the nonlinear sigma model, though not much of this has been proved. Unfortunately, to explain here what is known about (84) would require a lengthy explanation of the renormalization group, the $1/N$ expansion, factorizable S matrices, etc.

Now, the action function (67) and the integrand in (70) depend only on the conformal class of the metric tensor of Σ —in other words, they depend only on the complex structure of Σ . However, to define even formally an integration measure in (70) we needed to give Σ an actual Riemannian structure, not just a conformal class of Riemannian structures. Therefore, our eventual answer (79) is not a function on the moduli space of Riemann surfaces Σ —it is not invariant under a conformal rescaling of the metric of Σ . However, (79) can be shown to transform rather simply under a conformal rescaling of the metric tensor of Σ . It has been interpreted by Quillen as defining a metric on a certain holomorphic line bundle over moduli space [23]; this structure plays a significant role in the developments we hear about from Faltings in his lecture at the 1986 International Congress of Mathematicians.

The next major step in explaining what quantum field theory is as understood by physicists is to explain the relation between the Feynman path integrals which we have been describing and the older Hilbert space interpretation of quantum field theory. If the boundary of Σ is not empty, it consists of several disjoint circles S_i . Let us consider the case of a single circle S . Let $\Omega(S)$ be the space of continuous real-valued functions on S . As in our discussion of Σ , a Riemannian structure on S induces a Riemannian structure and therefore a measure on $\Omega(S)$. Given such a measure, we can speak of square integrable complex-valued

functions on $\Omega(S)$, that is, functions ψ with

$$\int_{\Omega(S)} |\psi|^2 < \infty. \quad (85)$$

Of course, it is necessary here to define precisely the measure on $\Omega(S)$, just as we needed zeta function (or some other) regularization in discussing integrals on $\Omega(\Sigma)$. While this can easily be made precise (at least in the case of free field theory), I will just proceed formally. Let us denote the Hilbert space of square integrable functions on $\Omega(S)$ as H_S . It is known as the Hilbert space of the quantum field theory that we are discussing.

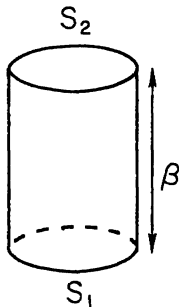


FIGURE 1. A cylinder with a standard, flat metric. The boundary consists of two components S_1 and S_2 of circumference 2π . The “height” of the cylinder is labeled as β .

Now I would like to define a certain linear operator on H_S known as the “Hamiltonian” of the theory in question. Let us consider a Feynman path integral (70) on a very special Riemann surface Σ with boundary—a cylinder with a standard, flat metric (see Figure 1). The boundary consists of two components, S_1 and S_2 . We will think of S_1 and S_2 as two copies of the “same” circle, displaced in the vertical direction through a distance (“proper time”) β . When desired we will feel free therefore to identify the function spaces $\Omega(S_1)$ and $\Omega(S_2)$ as well as the Hilbert spaces H_{S_1} and H_{S_2} . So for $\psi_1 \in H_{S_1}$ and $\psi_2 \in H_{S_2}$ we have an inner product

$$\langle \psi_2 | \psi_1 \rangle = \int_{\Omega(S)} \psi_2^* \psi_1. \quad (86)$$

A real-valued function f on the boundary of Σ is the same as a pair of functions f_1 and f_2 on the boundary components; we write $f = (f_2, f_1)$. We have already discussed in (70) the Feynman path integral on Σ with given boundary values:

$$Z_{(f_2, f_1)}(\Sigma) = \int_{\Omega_{(f_2, f_1)}} e^{-I}. \quad (87)$$

Now we will go a step further and define a bilinear functional on the Hilbert spaces H_{S_1} and H_{S_2} . Given $\psi_1 \in H_{S_1}$ and $\psi_2 \in H_{S_2}$, we define

$$R(\psi_2, \psi_1) = \int_{\Omega(S_2)} \psi_2^*(f_2) \int_{\Omega(S_1)} \psi_1(f_1) \cdot Z_{(f_2, f_1)}(\Sigma). \quad (88)$$

$R(\psi_2, \psi_1)$ is obviously linear in ψ_1 and antilinear in ψ_2 , so it has the general form

$$R(\psi_2, \psi_1) = \langle \psi_2 | T_\beta | \psi_1 \rangle \quad (89)$$

for some linear operator T_β . (Recall that β , indicated in Figure 1, is the "height" of the cylinder Σ .)

T_β is simply that linear operator on $\Omega(S)$ whose kernel is

$$T_\beta(f_2, f_1) = Z_{(f_2, f_1)}(\Sigma). \quad (90)$$

Now, suppose that, as in Figure 2, we glue together two cylinders Σ_1 and Σ_2 of thickness β_1 and β_2 along a common boundary component S to make a cylinder Σ of thickness $\beta = \beta_1 + \beta_2$.

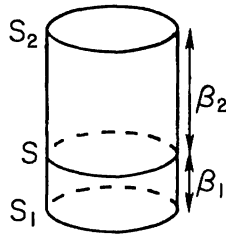


FIGURE 2. A cylinder of height $\beta_1 + \beta_2$, made by gluing together cylinders Σ_1 and Σ_2 of height β_1 and β_2 , respectively. The boundary of Σ_1 consists of the two circles S_1 and S , and the boundary of Σ_2 consists of the two circles S and S_2 .

The boundary of Σ_1 consists of two components, say S_1 and S , and the boundary of Σ_2 consists of two components, say S and S_2 . Suppose we fix real-valued functions f_1, f_2 , and f on S_1, S_2 and S , respectively, and look at the product

$$T_{\beta_2}(f_2, f)T_{\beta_1}(f, f_1) = Z_{(f_2, f)}(\Sigma_2)Z_{(f, f_1)}(\Sigma_1). \quad (91)$$

The product (91) is an integral over real-valued functions on Σ that have prescribed values on the three circles S_1, S_2 and S (functions, in other words, whose restriction to S_1, S_2 , or S coincides with f_1, f_2 , or f). If we do *not* wish to specify the values on S , we can avoid this by integrating (91) over $\Omega(S)$. We thus consider

$$\int_{\Omega(S)} T_{\beta_2}(f_2, f)T_{\beta_1}(f, f_1) = \int_{\Omega(S)} Z_{(f_2, f)}(\Sigma_2)Z_{(f, f_1)}(\Sigma_1). \quad (92)$$

Here we are integrating over real-valued functions on Σ whose values are specified only on $\partial\Sigma$ —in other words, we are considering our basic integral (70). Thus, (92) coincides with the kernel $T_\beta(f_2, f_1)$ associated according to (90) with integration over real-valued functions on the cylinder Σ of height $\beta = \beta_1 + \beta_2$:

$$T_{\beta_1 + \beta_2}(f_2, f_1) = \int_{\Omega(S)} T_{\beta_2}(f_2, f)T_{\beta_1}(f, f_1). \quad (93)$$

This equation is a semigroup law:

$$T_{\beta_1+\beta_2} = T_{\beta_2}T_{\beta_1}. \quad (94)$$

(94) means that

$$T_\beta = e^{-\beta H} \quad (95)$$

for some linear operator H . H is known as the “Hamiltonian” of the quantum field theory. In the free field theory that we have been discussing, an explicit formula for H is easily written. The reasoning by which we have given a Hilbert space interpretation to the Feynman path integral (70) and extracted a Hamiltonian is, however, far more general.

We would now like to compute the trace of the operator $e^{-\beta H} = T_\beta$. This is done, of course, by integrating the kernel $T_\beta(f_2, f_1)$ along the diagonal:

$$\text{Tre}^{-\beta H} = \int_{\Omega(S)} T_\beta(f, f). \quad (96)$$

In fact, (96) is simply an integral over real-valued functions on the cylinder Σ of Figure 1 with the boundary values on the two components identified but otherwise unrestricted. By identifying the boundary components of Σ one forms a Riemann surface (without boundary) of genus one, which we may call $\bar{\Sigma}$. (96) is simply the path integral of (70) carried out over $\bar{\Sigma}$:

$$\text{Tre}^{-\beta H} = \int_{\Omega(\bar{\Sigma})} e^{-I}. \quad (97)$$

Before discussing the significance of (97), we need a generalization. We return to the Hilbert space H_S associated with a circle S . The operation of rotating the circle by an angle θ obviously acts in a natural way on real-valued functions $\Omega(S)$ and therefore also on the Hilbert space H_S . We thus have a linear operator R_θ representing the action of the rotation on H_S . (We will use the same symbol R_θ to denote the action of this rotation on $\Omega(S)$.) We again obviously have a semigroup law, $R_{\theta_1+\theta_2} = R_{\theta_2}R_{\theta_1}$, so

$$R_\theta = e^{i\theta P} \quad (98)$$

where P is some linear operator, known as the momentum operator of the quantum field theory. Again, there is no difficulty in writing down explicit formulas for P . It is easy to see that it commutes with H . We now wish to generalize (97) and calculate the trace of the operator $T_{\beta,\theta} = e^{-\beta H + i\theta P}$. The kernel of this operator is simply

$$T_{\beta,\theta}(f_2, f_1) = T_\beta(f_2, R_\theta f_1). \quad (99)$$

Here $T_\beta(f_2, f_1)$ is the kernel of $e^{-\beta H}$, which we have discussed at length already, and the formula (99) holds because in $e^{-\beta H + i\theta P} = e^{-\beta H} \cdot e^{i\theta P}$, the effect of $e^{i\theta P}$ is just to replace f_1 with $R_\theta f_1$. Again constructing the trace by integrating the kernel along the diagonal, we have

$$\text{Tr} e^{-\beta H + i\theta P} = \int_{\Omega(S)} T_\beta(f, R_\theta f). \quad (100)$$

As in our previous discussion of the case $\theta = 0$, (100) has a simple interpretation. The right-hand side of (100) is, in the terminology of (70),

$$\int_{\Omega(S)} Z_{(f, R_\theta f)}(\Sigma) = \int_{\Omega(S)} \int_{\Omega_{(f, R_\theta f)}(\Sigma)} e^{-I}. \quad (101)$$

This is easily described in words. The right-hand side of (101) is a path integral on the cylinder of Figure 1. The integral ranges over all real-valued functions whose values on the boundary components S_1 and S_2 coincide after rotating S_1 through an angle θ . Thus, if we take our cylinder Σ of height β , and glue the two components together after rotating through an angle θ , we form a Riemann surface of genus one without boundary. We will call this surface $\Sigma(\beta, \theta)$. Our conclusion is

$$\text{Tr } e^{-\beta H + i\theta P} = \int_{\Omega(\Sigma(\beta, \theta))} e^{-I}. \quad (102)$$

(102) is a very general formula in quantum field theory. However, it is particularly interesting in the case in which the action functional I is conformally invariant—-independent of a conformal rescaling of the metric of Σ . Such theories are said to be conformal field theories. For example, free boson field theory is a conformal field theory. In the case of a conformal field theory, the right-hand side of (102) might appear at first sight to define a conformal invariant, depending only on the conformal structure on the surface $\Sigma(\beta, \theta)$. This is not quite true, because there is no conformally invariant way to define the measure in the integral over $\Omega(\Sigma)$. Nevertheless, it is possible to construct simple formulas for the deviation from conformal invariance of the right-hand side of (102). For mathematicians, these formulas involve the theory of the determinant line bundle; for physicists, they have involved the theory of anomalies.

It is well known that every Riemann surface of genus one is isomorphic to $\Sigma(\beta, \theta)$ for some values of β, θ in the range

$$0 \leq \beta < \infty, \quad 0 \leq \theta < 2\pi. \quad (103)$$

The correspondence is not one-to-one. If we introduce the complex variable $\tau = (\theta + i\beta)/2\pi$, then the group $\text{SL}(2, Z)$ acts on τ by

$$\tau \rightarrow \frac{a\tau + b}{c\tau + d}, \quad (104)$$

where a, b, c , and d are integers with $ad - bc = 1$. Riemann surfaces with values of τ that are related by (104) are isomorphic. In the theory of modular forms it is standard to introduce $q = e^{2\pi i\tau}$. If we define

$$H_\pm = (H \pm P)/2 \quad (105)$$

then (102) amounts to

$$\text{Tr } q^{H_+} \bar{q}^{H_-} = \int_{\Omega(\Sigma(\tau))} e^{-I}. \quad (106)$$

Conformal field theories in which $H_- = 0$ are said to be chiral theories. The free boson theory (67) is not chiral. The simplest example of a chiral theory is

the theory of free chiral fermions, but I will not explain this here. In a chiral theory, the right-hand side of (106) defines a holomorphic function of q or in other words “a modular function (perhaps only weakly modular).” In any case, in conformal field theories, whether chiral or not, one can obtain simple formulas for the transformation of (106) under conformal transformations, by using the theory of the determinant line bundle or the theory of anomalies [32, 9].

For example, in the theory of affine Lie algebras, it is known that the characters of the integrable highest weight modules transform in a simple way under $SL(2, Z)$. (See [18] for a survey.) At first sight this comes as something of a surprise; $SL(2, Z)$ does not enter the structure of these algebras in any obvious way; it is certainly not a group of automorphisms of them. However, it is known that the highest weight modules of affine Lie algebras have a natural description in terms of quantum field theory. For the level one modules of most of the affine Lie algebras this can be done using free bosons or free fermions [6, 20, 26, 7]. In general, arbitrary integrable highest weight modules of affine Lie algebras have quantum field theory realizations, albeit more complicated [33]. The quantum field theories in question are conformal field theories, so the appearance of $SL(2, Z)$ in the theory of affine Lie algebras is a special case of our assertion that $SL(2, Z)$ has a simple action on (106).

Most of the work on affine Lie algebras has been done in the context of the Hamiltonian formulation of quantum field theory. As we have just seen, the role of $SL(2, Z)$ becomes clear only in the path integral formulation. The discussion has made it clear that the Hamiltonian perspective is related to path integrals on a cylinder. In and of itself the Hamiltonian perspective seems self-contained. The path integral approach suggests the obvious generalization to other Riemann surfaces, which would scarcely occur to us if we think in Hamiltonian terms only. Given a quantum field theory whose path integral on a cylinder constructs a highest weight module of an affine Lie algebra, what is the mathematical significance of the same path integral on, say, a Riemann surface of g ? I cannot propose an answer to this question. One reason I hope that this question will attract interest is that a proper answer might well shed some light on string theory.

The path integrals that we have discussed until now are by no means the most general ones that are usually considered in quantum field theory. In fact, we have only considered very special cases of the usual structures. Usually one defines what are known as local operators. Consider, for instance, the free boson field theory which we have given as a simple example of quantum field theory. Let Σ be a Riemann surface. Let P be a point on Σ , and let x^i be local coordinates near P . In studying free boson field theory, we are integrating over real-valued functions

$$\phi: \Sigma \rightarrow R.$$

By a local operator at P we mean a functional O of ϕ which depends only on $\phi(P)$ and the derivatives of ϕ at P . More specifically, O is required to have only a polynomial dependence on the derivatives of $\phi(P)$. Thus, examples of local

operators would be

$$F(\phi(P)), \quad F^i(\phi(P)) \frac{\partial \phi}{\partial x^i}, \quad F^{ij}(\phi(P)) \frac{\partial^2 \phi}{\partial x^i \partial x^j}, \quad (107)$$

where F , F^i , F^{ij} are arbitrary real-valued functions of $\phi(P)$. (In string theory the case $F = e^{i\lambda\phi}$, $\lambda \in R$, is particularly important.) What do we do with such local operators? Let P_i be some points on Σ , and let O_i be local operators at P_i . Then we generalize (70) to consider

$$Z_f(O_i; \Sigma) = \int_{\Omega_f(\Sigma)} e^{-I} \prod_i O_i(P_i). \quad (108)$$

This would be described as a path integral on the surface Σ with insertion of the operators $O_i(P_i)$. One usually defines the "correlation function"

$$\langle O_1(P_1) O_2(P_2) \cdots O_n(P_n) \rangle = \frac{Z_f(O_i; \Sigma)}{Z_f(\Sigma)}. \quad (109)$$

For the significant choices of the O_i (for example, the operators (107) with $F = e^{i\lambda\phi}$), it is possible to work out quite explicit formulas for such correlation functions, which depend only on the Green function of the Laplacian Δ . I will not enter into this here.

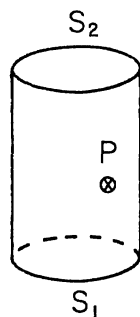


FIGURE 3. A path integral on a cylinder Σ with insertion of a local operator $O(P)$. The boundary of Σ again consists of the two circles S_1 and S_2 .

It remains to explain why the local functionals $O(P)$ are called local operators. Indeed, they correspond to operators in the Hamiltonian formulation of quantum field theory. Consider again a path integral on our familiar cylinder Σ , but now with an insertion of the operator $O(P)$ at the point P (see Figure 3). Let β_1 and β_2 be the distance from P to S_1 and S_2 , respectively. Generalizing (88), we consider the integral

$$R(O(P); \psi_2, \psi_1) = \int_{\Omega(S_2)} \psi_2^*(f_2) \int_{\Omega(S_1)} \psi_1(f_1) \cdot Z_{(f_2, f_1)}(O; \Sigma). \quad (110)$$

As in our previous case, linearity in the ψ_i ensures that (110) is of the form $\langle \psi_2 | U | \psi_1 \rangle$ for some operator U in the Hilbert space H_S of the quantum field theory. In fact, one defines a linear operator $\hat{O}(P)$ by

$$R(O(P); \psi_2, \psi_1) = \langle \psi_2 | e^{-\beta_2 H} \hat{O}(P) e^{-\beta_1 H} | \psi_1 \rangle. \quad (111)$$

This is a canonical correspondence between local functionals $O(P)$ that can be inserted in the path integral and local operators $\hat{O}(P)$ in the Hilbert space. (The precise sense in which the Hilbert space operator $\hat{O}(P)$ is "local" is something I will not enter into here.) Local operators $O(P)$ are crucial in physical applications of quantum field theory, and suitable local operators are the natural tools for describing the highest weight modules of affine Lie algebras.

There is in addition an important correspondence between operators and vectors in Hilbert space. Consider the simplest Riemann surface with boundary, namely, a disc D . Fix a point P in the interior of Σ . We have a Hilbert space H_S associated with real-valued functions on the boundary S of D . We also have local operators at P ; they form a vector space H_P . I will describe a natural map from H_P to H_S . Let $O(P)$ be a local operator at P . We wish to construct an element of H_S . Let f be a real-valued function on S , and let

$$\psi_O(f) = Z_f(O; D). \quad (112)$$

The correspondence $O \rightarrow \psi_O(f)$ gives the desired map $H_P \rightarrow H_S$. In conformal field theory, this correspondence is an isomorphism between H_S and H_P , and is of special importance.

One local operator of particular importance is the following. Under an infinitesimal change in the metric g_{ij} of Σ , the change in the action I of a quantum field theory is

$$\delta I = \int_{\Sigma} \delta g^{ij} T_{ij}, \quad (113)$$

where the symmetric tensor T_{ij} is known as the energy-momentum tensor. A conformal field theory is precisely one in which $g^{ij} T_{ij} = 0$. In a conformal field theory on a Riemann surface, T_{ij} has only two independent nonzero components, which transform as differentials of type (2,0) and (0,2), respectively. Let us call these T and \tilde{T} , respectively. Letting σ be an angular variable on our canonical circle S , we define

$$L_n = \int_S e^{in\sigma} T(\sigma), \quad \tilde{L}_n = \int_S e^{in\sigma} \tilde{T}(\sigma). \quad (114)$$

These can be shown to obey

$$\begin{aligned} [L_n, L_m] &= (m-n)L_{m+n} + c\delta_{m+n}(m^3 - m), \\ [\tilde{L}_n, \tilde{L}_m] &= (m-n)\tilde{L}_{m+n} + \tilde{c}\delta_{m+n}(m^3 - m), \\ [L_n, \tilde{L}_m] &= 0, \end{aligned} \quad (115)$$

where c and \tilde{c} are constants; δ_n is one for $n = 0$ and otherwise zero. Restricting ourselves to the L 's, (115) is the so-called Virasoro algebra, a central extension of the Lie algebra of diffeomorphisms of S . The latter algebra is generated by the vector fields

$$D_n = -ie^{in\sigma} d/d\sigma, \quad (116)$$

which are easily seen to obey $[D_n, D_m] = (m-n)D_{m+n}$; the first line in (115) is a central extension of (116). A representation of the Virasoro algebra is called

a highest-weight representation if L_0 is bounded above; a highest-weight vector is one annihilated by the L_n for $n > 0$. Conformal field theories lead to highest-weight representations of the Virasoro algebra, a special case of this being the assertion that the highest-weight representations of affine Lie algebras can be extended to the semidirect product with the Virasoro algebra.

The introduction to quantum field theory that I have given has been very sketchy to say the least. Quantum field theory is a rather rich and complex subject, and I have only explained a few generalities. Describing quantum field theory as a mathematical theory will become far easier and more natural once some of its characteristic mathematical applications begin to emerge. There are grounds for believing that—after sixty years—this time may be nearly at hand.

4. String theory. Finally, I would like to briefly describe what string theory is. String theory is a subject even more vast than those that we have considered previously, and my treatment will be even sketchier. Let us return to general relativity, described by the action functional

$$S = \frac{-1}{16\pi G} \int_M R, \quad (117)$$

with R being the Ricci scalar of a Riemannian metric g on a space-time manifold M . Treating (117) as a quantum theory would mean the following. Fixing the topology of M , let Λ be the space of metrics on M modulo diffeomorphisms. The quantization of general relativity would involve defining integrals such as

$$Z = \int_{\Lambda} e^{-S} \quad (118)$$

as well as generalizations with insertions of operators, modifications of boundary conditions if $\partial M \neq 0$, etc. While there is no rigorous theorem here, all the indications are that there is no satisfactory way to make sense of the integration measure in (118). This is one way of expressing the inconsistency between general relativity and quantum mechanics which was cited in the introduction as a central problem in theoretical physics.

Before discussing the string theory generalization of (118), let us discuss the perturbative expansion of general relativity. Let η_{ij} be the metric tensor of flat Minkowski space, and expand the metric tensor g_{ij} as

$$g_{ij} = \eta_{ij} + h_{ij}, \quad (119)$$

where h_{ij} is the metric disturbance. In an expansion near flat space, one can think of h_{ij} as taking values in a linear space Λ_0 . If the action function I is written in terms of h , the linear term vanishes because flat Minkowski space is a solution of the Einstein equations. The quadratic term is of the general form

$$S_2 = \frac{-1}{16\pi G} \int_M h \Delta_L h, \quad (120)$$

where Δ_L is a certain linear operator known as the Lichnerowicz Laplacian. The solutions of $\Delta_L h = 0$ are the linearized gravitational waves discussed in §1. The

cubic term is

$$S_3 = -\frac{1}{16\pi G}\Phi_3(h), \quad (121)$$

with $\Phi_3(h)$ being a complicated cubic expression, the details of which need not concern us. On substituting in (118) the expansion $S = S_2 + S_3 + \cdots$ of the action, one can try to develop a perturbation expansion, viewing h as a small quantity. The starting point, discarding all terms beyond S_2 , gives a Gaussian integral similar to those discussed in the last section. The correction terms can be computed, but in the case of general relativity, one runs into nonsensical infinite formulas. This is one major aspect of why we believe that general relativity does not make sense as a quantum theory.

In theories that do make sense as quantum theories, a major aspect of our understanding is the perturbative expansion analogous to the above. Such perturbative expansions are not beautiful. The beauty, if any, of a quantum field theory is to be found in the basic formulas, analogous to (117), not in the nitty-gritty of a perturbative expansion. Nevertheless, in the case of string theory, it is the perturbation expansion that we know. We do not know the basic formulas like (117), or the basic logical concepts that should play in string theory the role played by metrics, connections, and curvature in general relativity.

I will therefore explain in turn some of the ingredients which in string theory are analogous to h , Λ_0 , I_2 , and I_3 . First of all, the linear space Λ_0 is replaced by the Hilbert space of a certain quantum field theory. Our discussion of quantum field theory in the last section was rather formal, and we did not discuss the physical interpretation of the Riemann surface Σ or the quantum field ϕ . In traditional applications of quantum field theory, Σ plays the role of space-time, and ϕ is a field (analogous to the electromagnetic field) propagating in space-time. In string theory, the interpretation is reversed; the Riemann surface Σ is an auxiliary object, and ϕ is replaced by a map $\Phi: \Sigma \rightarrow M$, with M interpreted as space-time. Thus, let M be a flat manifold of dimension D , with standard coordinates X^i , $i = 1, \dots, D$. The map $\Phi: \Sigma \rightarrow M$ is specified by giving D real-valued functions X^i on Σ . The action functional is thus

$$I = \frac{1}{2\pi} \int_{\Sigma} \eta_{ij} dX^i \wedge *dX^j. \quad (122)$$

Clearly, (122) is invariant under D -dimensional Poincaré transformations of the X^i .

Now, in the linearized theory of general relativity, the main ingredients are the linear space Λ_0 of metric disturbances and the quadratic action functional S_2 on this space. What are the analogues of these in string theory? The analogue of Λ_0 is the Hilbert space H_S of the quantum field theory (122). The replacement of Λ_0 by H_S is really quite a drastic step, since H_S is a quite infinite space compared to Λ_0 . An element of Λ_0 is, concretely, a finite collection of functions

$$h_{ij}(x^k), \quad (123)$$

with x^k , $k = 1, \dots, D$, being the coordinates of the space-time manifold M . What is an element of H_S ? Letting σ be an angular parameter on the circle S ,

we defined in the last section the space Ω_S of maps from S to M . Such a map is concretely given by specifying real-valued functions $X^k(\sigma)$ for which we will make a Fourier expansion

$$X^k(\sigma) = x^k + \sum_{n \neq 0} e^{in\sigma} x_n^k. \quad (124)$$

We have separated out the zeroth Fourier component, which is known as the “center of mass of the string.” We would like to think of x^k in (124) as corresponding to the x^k in (123). Of course, x_n^k is the complex conjugate of x_{-n}^k . An element of H_S is a real-valued function on Ω_S , or concretely a function

$$\Phi(x^k; x_{\pm 1}^k, \dots). \quad (125)$$

Thus, Φ depends on an infinity of variables besides the center of mass coordinates x^k which appear already in (123). The sense in which string theory has field theory as an approximation is that if the x_n^k in (125) can be considered as small, then Φ reduces, in the first instance, to a function of the x^k just like the gravitational field or any other field normally considered in physics. To be somewhat more precise about this, suppose that the x_n^k for $n \neq 0$ are small enough that it is appropriate to make a Taylor series expansion about $x_n^k = 0$. The form of the expansion would be⁸

$$\Phi(x^k; x_n^k) = \phi(x^k) + x_1^j B_j(x^k) + x_1^i x_{-1}^j h'_{ij}(x^k) + \dots \quad (126)$$

In this expansion, the $\phi(x^k)$, $B_i(x^k)$, h'_{ij} , etc., are ordinary functions of the center of mass coordinates, that is, ordinary functions on M . The first point here is that we can think of Φ as an infinite collection of functions on space-time. General relativity is based on certain nonlinear partial differential equations for h_{ij} . In string theory, we will likewise write down nonlinear partial differential equations for Φ (or a somewhat larger set of variables). Concretely, a nonlinear equation for Φ can be understood as a system of coupled nonlinear equations for an infinite set of variables in space-time. The second major point about (126) is that there is a very definite sense in which Φ should be viewed as a generalization of (123). In (126), there appears the tensor field h'_{ij} , and this should be viewed as the analogue of h_{ij} . The nonlinear equation that one studies in string theory has the property that, when restricted to h'_{ij} , it reduces approximately to the Einstein equations obeyed by h_{ij} , on length scales large compared to the Planck length.

Having explained what is the string theoretic analogue of Λ_0 , the next step is to explain what is the analogue in string theory of the quadratic action functional S_2 of the linearized theory of general relativity. Roughly speaking, the analogue of S_2 is just

$$S_2(\Phi) = (\Phi | H | \Phi), \quad (127)$$

where H is the Hamiltonian of the quantum field theory (122). This isn't quite the right definition, but must be supplemented by a restriction to the highest-weight vectors of the Virasoro algebra. However, it will do for the moment.

⁸This is not precisely the right expansion to make, but it will do for our purposes here.

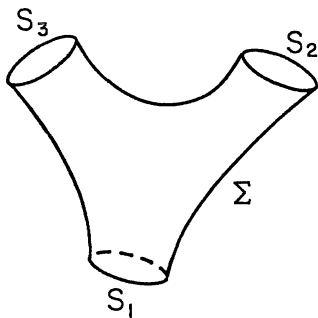


FIGURE 4. A Riemann surface with three boundary components.

What is the analogue in string theory of the nonlinear terms S_3 , etc., of general relativity? The right approach involves a simple elaboration of ideas from the last section. Consider a Riemann surface Σ whose boundary consists of three circles S_1 , S_2 , and S_3 , as in Figure 4. A simple variant of constructions in the last section is to consider a path integral on Σ in which the boundary condition on each boundary component is determined by Φ :

$$S_3(\Phi) = \int_{\Omega(S_1)} \Phi(f_1) \int_{\Omega(S_2)} \Phi(f_2) \int_{\Omega(S_3)} \Phi(f_3) Z_{(f_1, f_2, f_3)}(\Sigma). \quad (128)$$

This is then roughly the analogue of the cubic term in the Einstein action $S_3(h)$. It can obviously be generalized to a case with n boundary components.

There is, however, another language for thinking about this. At the end of the last section we noted that in conformal field theory there is a natural isomorphism $H_P \approx H_S$ between local operators $O(P)$ that can be inserted at a point P and vectors ψ_O in the Hilbert space H_S . The state Φ in (128) corresponds under this isomorphism to some operator V_Φ (called the vertex operator of Φ). Henceforth we will refer to V_Φ merely as V . According to (112), the relation between Φ and V is that if we introduce for $i = 1, 2, 3$ a disc D_i with boundary S_i , then

$$\Phi(f_i) = Z_{f_i}(V; D_i). \quad (129)$$

Suppose that we compactify the surface Σ of Figure 4 by gluing in the disc D_i along each boundary component S_i , for $i = 1, 2, 3$. Let us call the resulting closed surface $\bar{\Sigma}$.

We want to reexpress (128) as a path integral on $\bar{\Sigma}$. In (128), Φ has been used to define the boundary condition on each S_i . According to (129), defining the boundary condition on S_i by Φ is the same as gluing in the disc D_i with the operator V inserted on a point P_i of D_i . Thus, (128) is really equivalent to a path integral on $\bar{\Sigma}$ with insertions of the vertex operator V at the three points P_1 , P_2 , and P_3 :

$$S_3(\Phi) = Z(V(P_1), V(P_2), V(P_3); \bar{\Sigma}). \quad (130)$$

In (130) we have not been too particular about explaining which three points P_1 , P_2 , and P_3 have been chosen. There is a definite sense in which this does

not matter. The Riemann surface $\bar{\Sigma}$ is isomorphic to the Riemann sphere. It is well known that the group $SL(2, C)$ acts transitively on the configurations of three disjoint points on the Riemann sphere, so there is no conformal invariant associated with the choice of the P_i . If, however, we want to define not $S_3(\Phi)$ but an analogous object $S_n(\Phi)$ for $n > 3$, we face the question of choosing points P_i . The correct prescription, as articulated by Polyakov [22], is to integrate over the moduli of configurations of n points on $\bar{\Sigma}$:

$$S_n(\Phi) = \int_{\mathcal{M}_n} Z \left(\prod V(P_i); \bar{\Sigma} \right). \quad (131)$$

Here \mathcal{M}_n is the space of moduli of n disjoint points on the Riemann sphere. This integral is indicated in Figure 5 for the case of five points.

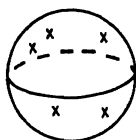


FIGURE 5. Integration over moduli of configurations of several points on the Riemann sphere.

In fact, (131) is not just the string-theoretic generalization of the Einstein action, but contains more information. (131) is really the string-theoretic analogue of the perturbation series constructed from (117) in defining the “tree approximation” to general relativity. The physical interpretation of (131) is that it gives the probability for scattering of n particles of type Φ . Unfortunately, to explain the latter remark would require a considerable enlargement of the brief introduction to quantum field theory in §3.

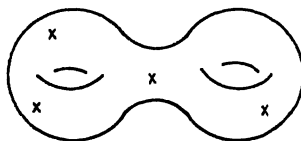


FIGURE 6. A path integral on a Riemann surface of genus two, with insertions of several vertex operators.

Now, in (131) $\bar{\Sigma}$ is a Riemann surface of genus zero, because we are led to genus zero in the course of trying to find a string theory analogue of (117). However, in the mathematical sense, (131) has a very natural generalization (Figure 6) with $\bar{\Sigma}$ replaced by a Riemann surface of genus greater than zero. This generalization is of central importance in string theory. I have remarked that if we try to interpret (117) as the Lagrangian of a quantum theory, we run

into severe difficulties. The attempt to calculate quantum corrections to the classical theory of general relativity gives rise to infinities. On the other hand, in string theory there is a meaningful and well-defined prescription for calculating quantum corrections to classical answers. One simply replaces $\bar{\Sigma}$ by a Riemann surface Σ_k of genus $k > 0$. Thus, if we want to calculate the probability of scattering of n particles of type Φ , then in string theory the classical answer, valid for $\hbar = 0$, is given by (131). If we want to calculate a quantum correction to (131) of order \hbar^k , then we calculate not (131) but

$$\int_{\mathcal{M}_{k,n}} Z \left(\prod V(P_i), \tilde{\Sigma} \right). \quad (132)$$

Here $\tilde{\Sigma}$ has genus k , and $\mathcal{M}_{k,n}$ is the moduli space of Riemann surfaces of genus k with n marked points. In this way, physicists have become interested in the moduli space of Riemann surfaces and in path integrals over this space. The integrals in (132) actually have remarkably beautiful properties, some of which were described by Yu. Manin in his 1986 International Congress of Mathematicians lecture, "Quantum Strings and Algebraic Curves." But even if (132) were not beautiful, the fact that it is free of the ultraviolet divergences that plague the analogous formulas in the quantum theory of general relativity would be enough to give it a far-reaching importance in physics.

There are many major gaps in this exposition. A much nicer description can be given if one considers not the space H_S but a certain highest-weight cohomology theory of the Virasoro algebra with values in H_S . The cohomology theory in question [3, 29, 19, 8] has been presented at this conference by I. Frenkel, and I will not enter into it here. Also, I have avoided the question of what vertex operators V we are using in (131) and (132). These formulas are actually limited to vertex operators V that transform on the Riemann surface like differential forms of type (1,1); they correspond, under our canonical correspondence between operators and vectors in H_S , to highest-weight vectors of the Virasoro algebra. The highest-weight cohomology theory that I just mentioned is the proper framework for formulating the quadratic action $S_2(\Phi)$ [27, 1, 28] and is also very useful in what little we can say about the nonlinear theory which has the perturbative expansion we have discussed [34, 15, 21].

I have tried to make it plausible that path integrals on Riemann surfaces can be used to formulate a generalization of general relativity. What is more, the resulting generalization is (especially in its supersymmetric forms) free of the ailments that plague quantum general relativity. If the logic has seemed a bit thin, it is at least in part because almost all we know in string theory is a trial and error construction of a perturbative expansion. (131) and (132) are probably the most beautiful formulas that we now know of in string theory, yet these formulas are merely a perturbative expansion (in powers of Φ and \hbar) of some underlying structure. Uncovering that structure is a vital problem if ever there was one.

REFERENCES

1. T. Banks and M. Peskin, *Gauge invariance of string fields*, Nuclear Phys. B **264** (1986), 513.
2. J.-P. Bismut and D. Freed, *Analysis of elliptic families*, Preprint, 1985.
3. C. Becchi, A. Rouet, and R. Stora, Ann. Physics **98** (1974), 287.
4. P. Candelas, G. Horowitz, A. Strominger, and E. Witten, Nuclear Phys. B **258** (1985), 46.
5. T. Appelquist, A. Chodos, P. G. O. Freund (Editors), *Modern Kaluza-Klein theory and its applications*, Benjamin-Cummings (to appear).
6. A. Feingold and J. Lepowsky, Adv. in Math. **29** (1978), 271.
7. I. Frenkel and V. Kac, Invent Math. **62** (1980), 23.
8. I. Frenkel, H. Garland, and G. Zuckerman, Preprint, Yale Univ., New Haven, Conn., 1986.
9. D. Freed, MIT Preprint, 1986.
10. J. Glimm and A. Jaffe, *Quantum physics: A functional integral point of view*, Springer-Verlag, 1981.
11. M. B. Green and D. J. Gross (Editors), *Unified string theories*, World Scientific, Singapore, 1986.
12. M. B. Green and J. H. Schwarz, Phys. Lett. **54B** (1985), 502.
13. M. B. Green, J. H. Schwarz, and E. Witten, *Superstring theory*, Cambridge Univ. Press, New York, 1987.
14. D. J. Gross, J. A. Harvey, E. Martinec, and R. Rohm, Phys. Rev. Lett. **54** (1984), 502.
15. H. Hata, K. Itoh, T. Kugo, H. Kunitomo, and K. Ogawa, *Covariant string field theory*, Preprint, Kyoto Univ., 1986.
16. C. Itzykson and J.-B. Zuber, *Quantum field theory*, McGraw-Hill, New York, 1980.
17. M. Jacob (Editor), *Dual models*, North-Holland, Amsterdam, 1974.
18. V. Kac, *Infinite dimensional Lie algebras*, Birkhauser, 1983.
19. M. Kato and K. Ogawa, Nuclear Phys. B **212** (1983), 443.
20. J. Lepowsky and R. L. Wilson, Comm. Math. Phys. **62** (1978), 43.
21. A. Neveu and P. C. West, Phys. Lett. B **168** (1986), 192.
22. A. M. Polyakov, Phys. Lett. B **203** (1981), 207.
23. D. Quillen, *Determinants of Cauchy-Riemann operators on a Riemann surface*, Preprint, Inst. Hautes Études Sci.
24. J. Scherk and J. H. Schwarz, Phys. Lett. B **52** (1974), 374.
25. J. H. Schwarz (Editor), *Superstrings*, World Scientific, Singapore, 1986.
26. G. Segal, Comm. Math. Phys. **80** (1981), 301.
27. W. Siegel, Phys. Lett. B **151** (1985) 391, 396.
28. W. Siegel and B. Zweibach, Nuclear Phys. B **263** (1986), 105.
29. I. V. Tyupin, Lebedev preprint FIAN, No. 39, 1975.
30. G. Veneziano, Nuovo Cimento A **57** (1986), 190.
31. E. Witten, *Fermion quantum numbers in Kaluza-Klein theory*, Shelter Island II: Proceedings of the 1983 Shelter Island Conference on Quantum Field Theory and the Fundamental Problems of Physics, N. Khuri et al., editors, M.I.T. Press, Cambridge, Mass., 1985.
32. —, *Global anomalies in string theory*, Anomalies, Geometry, and Topology, A. White and W. Bardeen, editors, World Scientific, Singapore, 1985.
33. —, Comm. Math. Phys. **92** (1984), 495.
34. —, Nuclear Phys. B **268** (1986), 253.
35. A. Zee (Editor), *Unity of forces in the universe*, World Scientific, Singapore, 1982.

PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08544, USA

INVITED FORTY-FIVE MINUTE ADDRESSES
AT THE SECTION MEETINGS

The Complexity of Antidifferentiation, Denjoy Totalization, and Hyperarithmetic Reals

ALEXANDER S. KECHRIS

1. The problem of the primitive. We consider real functions on the interval $[0, 1]$. Denote by Δ the set of derivatives; i.e.,

$$\begin{aligned}\Delta &= \{f: f \text{ is a derivative}\} \\ &= \{f: \exists F: [0, 1] \rightarrow \mathbf{R} \text{ (} F \text{ is differentiable and } f = F')\}.\end{aligned}$$

If $f \in \Delta$, any F with $F' = f$ is a *primitive* of f and is uniquely determined up to a constant. To normalize, we denote by $F(x) = \int_0^x f$ the unique primitive of f with $F(0) = 0$. This is the original Newtonian concept of integration as the inverse operation of differentiation, i.e., antidifferentiation.

We will be concerned here with some definability aspects of the

CLASSICAL PROBLEM OF THE PRIMITIVE. Reconstruct the primitive F of a given derivative f .

This goes back to Newton and Leibniz and has been considered over the years in a different light as the concept of function has evolved (see [L2, P, S]). In “modern times” this problem was solved by Cauchy for continuous f (Cauchy definition of the integral) and Riemann for Riemann-integrable f (Riemann integration). Various generalizations were developed in the last half of the nineteenth century until Lebesgue introduced in his thesis (“Intégrale, Longueur, Aire,” 1902) his concept of integral. As explained in his book *Leçons sur l'Intégration et la Recherche des Fonctions Primitives* [L1], published in 1904, his primary motivation was the solution of the problem of the primitive in the general case of an arbitrary derivative. His Lebesgue integral made a major breakthrough here and totally resolved the problem in the case of a bounded derivative, more generally a Lebesgue integrable one. However, Lebesgue regretfully pointed out that he was unable to solve the problem completely for nonintegrable f . He considered this a major open problem that awaited solution. Such a solution was finally achieved in 1912 by A. Denjoy. The second edition of Lebesgue's aforementioned book [L2] published in 1928, contains an extensive treatment of Denjoy's work.

Research partially supported by National Science Foundation Grant DMS-8416349.

2. Denjoy totalization. In his solution to the problem of the primitive, Denjoy developed a concept of integration called *Denjoy totalization*. It consists of a transfinite iteration of Lebesgue integrations, computations of limits of sequences, and summations of series. (It is worth noting that transfinite iteration has also been used earlier in the context of Riemann as well as Lebesgue integration—even by Lebesgue himself.) We will briefly review now the concept of Denjoy totalization.

We begin with the following result of Lebesgue.

PROPOSITION 2.1. *Let $E \subseteq [a, b]$ be closed and $\{(a_i, b_i)\}$ the intervals contiguous to E in $[a, b]$ (i.e., the components of its complement). Let F be differentiable on $[a, b]$ and assume F' is Lebesgue integrable on E and $\sum |F(b_i) - F(a_i)| < \infty$. Then*

$$F(b) - F(a) = \int_E F'(x) dx + \sum [F(b_i) - F(a_i)].$$

This motivates the following definitions.

DEFINITION 2.2. Let E be a closed set and f a Lebesgue measurable function. We say that a point $x \in E$ is a point of *nonsummability* of f on E if f is not Lebesgue integrable in every $I \cap E$, I an open interval containing x .

DEFINITION 2.3. Let F be a continuous function on $[a, b]$, let $E \subseteq [a, b]$ be closed, and let $\{(a_i, b_i)\}$ be the contiguous intervals of E in $[a, b]$. A point $x \in E$ is called a point of *divergence* of F on E if $\sum_I |F(b_i) - F(a_i)| = \infty$ for all open intervals I containing x , where \sum_I indicates that we include only the (a_i, b_i) contained in I .

One has now the following fact.

PROPOSITION 2.4. *If F is a differentiable function on $[a, b]$ and $E \subset [a, b]$ is closed, then the points of nonsummability of F' on E and the points of divergence of F on E form a closed nowhere dense set in E .*

We can describe now the process of Denjoy totalization:

Given $f(= F')$ we reconstruct the primitive F —or actually the differences $F(b) - F(a)$ for all $a < b$ in $[0, 1]$ —by transfinite induction as follows. We construct a decreasing transfinite sequence $E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots \supseteq E_\alpha \supseteq \cdots$ of closed subsets of $[0, 1]$, each of which is nowhere dense in its predecessor (at successor stages), and simultaneously for each α the differences $F(b) - F(a)$ for (a, b) disjoint from E_α . Since for some countable α_0 we must have $E_{\alpha_0} = \emptyset$, it follows that by stage α_0 we have constructed $F(b) - F(a)$ for all $a < b$ in $[0, 1]$; i.e., we have reconstructed the primitive of f .

To start with, we let

$$E_1 = \{x \in [0, 1] : x \text{ is a point of nonsummability of } f \text{ on } [0, 1]\}.$$

Let $\{(a_i^1, b_i^1)\}$ be the contiguous intervals of E_1 in $[0, 1]$. Then $F(b) - F(a)$ is computed by Lebesgue integration for all $a < b$ with $[a, b] \cap E_1 = \emptyset$, and thus by taking limits (since F is continuous) we can compute $F(b_i^1) - F(a_i^1)$ as well.

Now let

$$E_2 = \{x \in E_1 : x \text{ is a point of nonsummability of } f \text{ on } E_1 \\ \text{or } x \text{ is a point of divergence of } F \text{ on } E_1\}.$$

Note that this makes sense since we know $F(b_i^1) - F(a_i^1)$. Then again, let $\{(a_i^2, b_i^2)\}$ be the intervals contiguous to E_2 in $[0, 1]$. Using Proposition 2.1 we can now compute $F(b) - F(a)$ for all $[a, b]$ disjoint from E_2 (letting $E = E_1 \cap [a, b]$ there), by the formula

$$F(b) - F(a) = \int_E f(x) dx + \sum_{[a,b]} [F(b_i^1) - F(a_i^1)].$$

Then we compute $F(b_i^2) - F(a_i^2)$ by taking limits as before.

We proceed this way by transfinite induction, taking intersections at limit ordinals.

Of course the concept of arbitrary (countable) ordinals is essential in Denjoy totalization, and Denjoy as well as Lebesgue considered that a major application of the ideas of Cantor's transfinite set theory to analysis and therefore, in their view, an important justification of this theory. (Denjoy has, of course, repeatedly emphasized the relevance of transfinite set theory in analysis (see [D1] and [D2]).

Although Denjoy totalization employs only countably many ordinals for each given derivative f , Denjoy constructed examples of f 's for which his process takes arbitrarily long countably many steps. In other words, the totality of all countable ordinals is necessary in his totalization.

This problem was therefore raised: to what extent is the transfinite induction and the use of the totality of countable ordinals necessary in the problem of reconstructing the primitive (as opposed to its use in a particular process for doing that). See, for example, [P, pp. 170–171], which refers to Lusin's Thesis 1915 as one of the first places where this question has been discussed. Other definitions of integrals (avoiding any use of ordinals) have been proposed, such as the Perron integral, the Kurzweil-Henstock integral, etc. (see [B]). These can be used to recover the primitive of any derivative, but they are hardly constructive in any sense.

3. The complexity of antidifferentiation. The preceding discussion brings us to some recent results which address these problems from the point of view of logic and definability theory—more particularly, in this case, *descriptive set theory*. The idea is to classify the complexity of the operation of antidifferentiation $(*)f \mapsto \int f$. If one can show that this is sufficiently complex (not Borel as we explain in a moment), then this indicates that there is no simple constructive notion of integral (some kind of “super” Lebesgue integration) which is sufficient to invert any derivative. Thus, this shows the necessity of the use of transfinite induction over all the countable ordinals in any constructive such process.

First we have to make precise in what sense we will express the complexity of $(*)$, and this amounts to making clear in what sense a derivative f is considered

as “given.” Since every derivative is a Baire class 1 function, the most reasonable way to consider a derivative as “given” is via a “code” of it as a Baire class 1 function, i.e., in terms of a sequence of continuous functions pointwise converging to it. (One could argue that giving such a sequence gives in some sense too much information about f . But since the point of the results below is that one cannot define simply the primitive of f , *even with this extra information given*, this is even better.)

So let $C[0, 1]$ denote the Polish space of continuous functions on $[0, 1]$ with the uniform metric and let $C[0, 1]^{\mathbb{N}}$ be the Polish space of infinite sequences of continuous functions with the product topology. Let

$$CN = \{\bar{f} \in C[0, 1]^{\mathbb{N}} : \forall x (\{f_n(x)\} \text{ converges})\}$$

be the class of pointwise converging sequences of continuous functions, and for $\bar{f} \in CN$ let $f = \lim \bar{f}$ be defined by $f(x) = \lim f_n(x)$ for all $x \in [0, 1]$. It is not hard to verify that CN is a complete \prod_1^1 subset of $C[0, 1]^{\mathbb{N}}$, so it is not Borel.

Now let

$$\bar{\Delta} = \{\bar{f} \in C[0, 1]^{\mathbb{N}} : \bar{f} \in CN \text{ and } \lim \bar{f} \in \Delta\}$$

be the set of codes of derivatives. The first result here was proven by M. Ajtai several years ago and provides an upper bound for the complexity of $\bar{\Delta}$.

THEOREM 3.1 (AJTAI, UNPUBLISHED). *The set $\bar{\Delta}$ is \prod_1^1 .*

The following lower bound completes the classification.

THEOREM 3.2 [K]. *The set $\bar{\Delta}$ is not \sum_1^1 ; i.e., there is no \sum_1^1 set $S \subseteq CN$ such that for $\bar{f} \in CN$, $\bar{f} \in S \Leftrightarrow \bar{f} \in \bar{\Delta}$. The same holds even for $\overline{b_1\Delta} = \{\bar{f} \in CN : \lim \bar{f} \text{ is a derivative of absolute value } \leq 1\}$.*

This computes the complexity of the domain of the operation of antidifferentiation. For the operation itself one first has the following upper bound.

THEOREM 3.3 (AJTAI, UNPUBLISHED). *The operation of antidifferentiation is Δ_1^1 . More precisely there are \sum_1^1 , \prod_1^1 relations $S, P \subseteq C[0, 1]^{\mathbb{N}} \times \mathbb{R}^2$ such that for $\bar{f} \in CN$, with $f = \lim \bar{f} \in \Delta$ and all $x \in [0, 1]$, $y \in \mathbb{R}$, we have*

$$y < \int_0^x f \Leftrightarrow S(\bar{f}, x, y) \Leftrightarrow P(\bar{f}, x, y).$$

Notice that by standard effective descriptive set theory Theorem 3.3 \Rightarrow Theorem 3.1. Ajtai's proof of Theorem 3.1 and 3.3 used nonstandard models and Denjoy totalization (oral communication). An alternative way to prove Theorem 3.3 is to show that the transfinite process in the Denjoy totalization for f terminates at some ordinal recursive in \bar{f} (where $\lim \bar{f} = f$). This approach uses also a crucial boundedness argument due to H. Woodin.

On the other hand, it was recently established that antidifferentiation is not Borel.

THEOREM 3.4 (DOUGHERTY-KECHRIS [DK]). *The operation of antidifferentiation is not Borel. More precisely, there is no Borel set $B \subseteq C[0, 1]^{\mathbb{N}}$ such that for $\bar{f} \in CN$, with $f = \lim \bar{f} \in \Delta$, $\bar{f} \in B \Leftrightarrow \int_0^1 f > 0$.*

This shows that the totality of countable ordinals is necessary in any constructive process for recovering the primitive, since no “simple analytical” operation suffices for this purpose. We take here as a necessary characteristic of such an operation that it is Borel (in the codes). This is clearly the case for Riemann or Lebesgue integration, taking of limits of sequences, summation of series, etc. So one can argue that arbitrary countable ordinals are intrinsically connected with the operation of recovering the primitive itself, and not only with a particular process of reconstruction, like Denjoy’s.

4. New characterizations of the hyperarithmetical reals. What is behind the preceding results is actually a new characterization of the hyperarithmetical reals. A sequence $\bar{f} \in C[0, 1]^{\mathbb{N}}$ is called *recursive* if it is a recursive sequence of recursive functions. Then Theorem 3.3 can be rephrased as follows.

THEOREM 4.1 (AJTAI, UNPUBLISHED). *Let $\bar{f} \in C[0, 1]^{\mathbb{N}}$ be recursive, $\bar{f} \in CN$, $f = \lim \bar{f} \in \Delta$. Then the primitive $\int f$ is Δ_1^1 . (Similarly relativized to any given real.)*

Note that $F = \int f$ is by its definition a \prod_1^1 singleton (if $f = \lim \bar{f}$, \bar{f} recursive). That it is actually Δ_1^1 is based ultimately on the Denjoy totalization, so this gives a definability consequence of this process, which clearly quantifies its constructive aspects.

Recall that a real $x \in \mathbb{R}$ is *hyperarithmetical* (or Δ_1^1) if $\{r \in \mathbb{Q}: r < x\}$ is hyperarithmetical. Call also a derivative $f \in \Delta$ *recursive* if there is recursive $\bar{f} \in CN$, with $f = \lim \bar{f}$. Then one has immediately from Theorem 4.1 that for a recursive derivative f , $\int_0^1 f$ is a hyperarithmetical real (and similarly relativized to any given real). The main result of [DK] is the converse to this.

THEOREM 4.2 (DOUGHERTY-KECHRIS [DK]). *Let $x \in \mathbb{R}$. Then the following are equivalent:*

- (i) x is hyperarithmetical,
 - (ii) $x = \int_0^1 f$, for some recursive derivative f .
- (Similarly relativized to any given real.)

The proof of this theorem involves effective transfinite induction based on the Recursion Theorem.

One can further combine Theorem 4.2 with Matyasevich’s Theorem to obtain formulas and an equivalent characterization of the hyperarithmetical reals involving only classical notions of analysis.

THEOREM 4.3 (DOUGHERTY-KECHRIS [DK]). *The hyperarithmetic reals are exactly those of the form $\int_0^1 f$, where $f(x)$ is a derivative given by*

$$f(x) = \sum_{n=0}^{\infty} a_n \frac{\max\left(0, 1 - \left| \left(n - [\sqrt{n}]^2 + 1 \right) x - \left([\sqrt{n}] - ([\sqrt{n}]^2) \right) \right| \right)}{n - [\sqrt{n}]^2 + 1},$$

with

$$a_n = (-1)^n \sum_{0 \leq m_1, \dots, m_N \leq 2^{2^N}} \max(0, 1 - Q(n, m_1 \cdots m_N)^2),$$

and Q an exponential polynomial with coefficients in \mathbf{Z} .

Another version of this kind of result can be stated as follows: call a function *analytically expressible* if it can be expressed by an explicit formula involving elementary functions and $\sum_{n=0}^{\infty}$. Then one can show that the hyperarithmetic reals are exactly the reals of the form $\int_0^1 \varphi$, where φ is an analytically expressible derivative. This demonstrates clearly the “definability gap” between a function and its derivative. One can have a derivative given by a simple analytical formula, while its primitive is immensely complex.

5. Some open problems. (A) The first problem is related to definability aspects of so-called “descriptive definitions of integrals” (see [S, Chapters VII, VIII]). These are essentially implicit definitions like the original one of the primitive. For example, the Lebesgue integral F of an integrable function f can be defined as the unique (up to a constant) F which is such that (i) F is absolutely continuous, and (ii) $F'(x) = f(x)$ for almost all x . By replacing, in (i), absolute continuity by more general conditions, one can obtain descriptive definitions of integrals inverting any derivative. The question is whether these conditions can possibly be Borel. (Note that (ii) is clearly Borel and so is the condition of absolute continuity.) This leads to the following

PROBLEM. Is there a Borel relation $B \subseteq C[0, 1]^{\mathbf{N}} \times C[0, 1]$ such that if $\bar{f} \in C^{\mathbf{N}}$, $f = \lim \bar{f} \in \Delta$, and $F \in C[0, 1]$, then $F = \int f \Leftrightarrow (\bar{f}, F) \in B$. We conjecture that the answer is no. (This would be stronger than Theorem 3.4.)

(B) An interesting but a bit vague problem is to come up with simpler formulas in Theorem 4.3. Another more precise question is the following. If in Theorem 4.3 we expand each summand of f in a Fourier series, then we obtain a formula for the hyperarithmetic reals of the form $\int_0^1 \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} a_{nm} \cos m(x - b_n)$, with explicitly given a_{nm} , b_n . Can we express them simply as

$$\int_0^1 \sum_{n=0}^{\infty} a_n \cos n(x - b_n),$$

with recursive $\{a_n\}$, $\{b_n\}$?

(C) The final problem is related to another important result of Denjoy. If a 2π -periodic function $f(x)$ has a trigonometric expansion

$$f(x) = \sum_{n=0}^{\infty} (a_n \cos nx + b_n \sin nx),$$

then the coefficients $\{a_n\}, \{b_n\}$ are uniquely determined by Cantor's Uniqueness Theorem. If f is Lebesgue integrable, the a_n, b_n can be computed by Fourier's formula. But how does one find a_n, b_n from f in the general case? This problem was solved by Denjoy in 1921 (see [D1]) by an extremely complicated procedure, again involving transfinite induction. From the definability point of view this leads to the following two questions:

PROBLEM. Classify the complexity of $T = \{\bar{f} \in CN: \lim \bar{f} \text{ admits a trigonometric expansion}\}$.

PROBLEM. Classify the complexity of the operation $\bar{f} \mapsto \{a_n\}, \{b_n\}$, where $\bar{f} \in T, \lim \bar{f} = f = \sum (a_n \cos nx + b_n \sin nx)$.

From its definition T is Σ_2^1 and this operation is Δ_2^1 on T . However, it appears that Denjoy's constructive process for recovering a_n, b_n ought to lead to a substantial lowering of this complexity, perhaps at the level of Σ_1^1 or Π_1^1 nonmonotone inductive definitions. If this is correct and if one can show corresponding lower bounds, this would lead to a classification of a natural and basic concept of analysis which falls between two levels of the projective hierarchy. This certainly would be a very interesting phenomenon. Also, if it is indeed true that the complexity of computing the trigonometric expansion is quite high above Borel, this would give a nice definability "explanation" of the considerable difficulty of Denjoy's procedure. At this point, however, this is only speculation, since it is not even known whether $\bar{f} \mapsto \{a_n\}, \{b_n\}$ is Borel or not.

REFERENCES

- [B] P. Bullen, *Non-absolute integrals: A survey*, Real Anal. Exchange, 5 (1979-80), 195-259.
- [D1] A. Denjoy, *Leçons sur le calcul des coefficients d'une série trigonométrique*, I-IV, Gauthier-Villars, Paris, 1941-49.
- [D2] —, *L'énumération transfinie*, Gauthier-Villars, Paris, 1946.
- [DK] R. Dougherty and A. Kechris, *The complexity of antidifferentiation*, Nov. 1985; Addendum, *Analytical expressions for the hyperarithmetical reals*, Feb. 1986; mimeographed circulated notes.
- [K] A. Kechris, *Classifying the complexity of the set of derivatives*, July 1985, mimeographed circulated notes.
- [L1] H. Lebesgue, *Leçons sur l'intégration et la recherche des fonctions primitives*, Gauthier-Villars, Paris, 1904.
- [L2] —, *ibid.*, 2eme éd., Paris, Gauthier-Villars, 1928.
- [P] I. Pesin, *Classical and modern integration theories*, Academic Press, New York, 1970.
- [S] S. Saks, *Theory of the integral*, 2nd rev. ed., Dover, 1964.

CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA 91125, USA

Homogeneous Structures

A. H. LACHLAN

Our purpose is to survey what is known about homogeneous structures over a finite relational language. We will sketch the main results obtained so far, and, by making some conjectures, will identify the most serious gaps in our knowledge. The situation can be summarized as follows: Finite homogeneous structures are well understood. Stable homogeneous structures turn out to be just the unions of chains of finite ones. Thus, understanding stable homogeneous structures goes hand in hand with understanding finite ones. Beyond this, some special cases have been investigated successfully, but almost no general results have been obtained.

The results of the work on special cases consisting of exhaustive lists of the homogeneous structures of various particular kinds, e.g., graphs, are described in §2. In §3 we survey the theory of stable homogeneous structures, and in §4 we speculate on what might be true in general.

1. Preliminaries. Let L be a finite relational language and \mathcal{M} an L -structure. Then M denotes the universe of \mathcal{M} and $\text{Th}(\mathcal{M})$ its first-order theory. \mathcal{M} is *imprimitive* if there is a nontrivial equivalence relation on M 0-definable in \mathcal{M} (i.e., definable by a formula with no parameters). Otherwise, \mathcal{M} is *primitive*.

The L -structure \mathcal{M} is *homogeneous* if it is countable and any isomorphism between finite substructures extends to an automorphism of \mathcal{M} . The class of all homogeneous L -structures is denoted $\text{Hom}(L)$.

Let $\mathbf{F}(L)$ denote the class of all finite L -structures. A subclass $\mathbf{U} \subseteq \mathbf{F}(L)$ is called *universal* if it is closed under isomorphism and substructures. \mathcal{M} is a *constraint* of \mathbf{U} if all its proper substructures lie in \mathbf{U} but \mathbf{U} itself does not. The closure \mathbf{U}^c of \mathbf{U} is the class of all L -structures which are unions of ascending chains of members of \mathbf{U} . A universal class \mathbf{U} is *finitely constrained* or *strictly universal* if its class of constraints is finite modulo isomorphism. A class $\mathbf{A} \subseteq \mathbf{F}(L)$ has the *amalgamation property* (AP) if \mathbf{A} is universal and for all $\mathcal{M}, \mathcal{M}_0, \mathcal{M}_1$ in \mathbf{A} and embeddings $F_i: \mathcal{M} \rightarrow \mathcal{M}_i$ ($i = 0, 1$) there exist \mathcal{N} in \mathbf{A} and embeddings $G_i: \mathcal{M}_i \rightarrow \mathcal{N}$ such that $G_0 F_0 = G_1 F_1$. Also, $\mathbf{A} \subseteq \mathbf{F}(L)$ has the *joint embedding property* (JEP) if for all $\mathcal{M}_0, \mathcal{M}_1$ in \mathbf{A} there exists \mathcal{N} in \mathbf{A} in which both \mathcal{M}_0

and \mathcal{M}_1 are embeddable. $\mathbf{A} \subseteq \mathbf{F}(L)$ is an *amalgamation class* if \mathbf{A} is nonempty and has both AP and JEP.

The language L is called *binary* if all the relation symbols are either unary or binary. An important observation is:

PROPOSITION. *Let L be binary and a finite set \mathbf{B} of finite L -structures be given. We can check effectively whether the universal class $\mathbf{U} \subseteq \mathbf{F}(L)$ constrained by the closure of \mathbf{B} under isomorphism is an amalgamation class.*

This proposition holds because in checking AP we need only look at one-point amalgamations, i.e., at amalgamations in which \mathcal{M}_0 and \mathcal{M}_1 have only one more element than \mathcal{M} , and then in checking JEP we need only look at one-point structures \mathcal{M}_0 and \mathcal{M}_1 .

With each L -structure \mathcal{M} we associate $\mathbf{S}(\mathcal{M})$, the class of all finite L -structures which are embeddable in \mathcal{M} . The close relationship between homogeneity, amalgamation classes, and admitting elimination of quantifiers was observed by Fraïssé [F] some thirty years ago and is summarized in the following theorem.


THEOREM 1. *Let L be a finite relational language.*

- (1) *For an L -structure \mathcal{M} , \mathcal{M} is homogeneous iff $\text{Th}(\mathcal{M})$ admits elimination of quantifiers.*
- (2) *If $\mathbf{A} \subseteq \mathbf{F}(L)$ is an amalgamation class, then there exists $\mathcal{M} \in \text{Hom}(L)$, unique up to isomorphism, such that $\mathbf{A} = \mathbf{S}(\mathcal{M})$.*
- (3) *If $\mathcal{M} \in \text{Hom}(L)$, then $\mathbf{S}(\mathcal{M})$ is an amalgamation class.*
- (4) *$\text{Hom}(L)$ is an elementary class.*

From (2) and (3) we see that the study of homogeneous structures is the same as the study of amalgamation classes. From (1) we deduce that if $\mathcal{M} \in \text{Hom}(L)$, then $\text{Th}(\mathcal{M})$ is \aleph_0 -categorical.

If $\text{Th}(\mathcal{M})$ is \aleph_0 -categorical, then for each $n < \omega$, the subsets of M^n invariant under $\text{Aut}(\mathcal{M})$ are the same as those definable without parameters in \mathcal{M} . Thus, instead of studying \mathcal{M} , we can study the permutation structure $(M, \text{Aut}(\mathcal{M}))$. This seemingly trivial remark turns out to be extremely useful, particularly when dealing with finite structures, because it makes available deep results from the theory of permutation groups. Let X be a countable set and G a subgroup of $\text{Sym}(X)$, i.e., let (X, G) be a permutation group of countable degree. (X, G) is then a *permutation structure* if G is complete in the topology of pointwise convergence, and for each $n < \omega$, X^n has finitely many orbits under G . The permutation structures are just the pairs $(M, \text{Aut}(\mathcal{M}))$ which arise from countable \aleph_0 -categorical structures. For $1 \leq k < \omega$, call (X, G) *k-ary* if for $n < \omega$ and tuples $\bar{a}, \bar{b} \in X^n$ not conjugate under G , there are corresponding subsequences \bar{a}', \bar{b}' of \bar{a}, \bar{b} having length at most k which are also not conjugate under G . If k is the maximum arity of the relation symbols of L , then the permutation structures arising from structures in $\text{Hom}(L)$ are *k-ary*. Conversely, every *k-ary* permutation structure arises in this way.

2. Examples. Let L_0 be the language with one binary relation symbol R . Most of the special cases mentioned in the introduction concern the intersection of $\text{Hom}(L_0)$ with some universal class of L_0 -structures. Below we shall give tables listing the homogeneous graphs, partial orders, tournaments, etc. In each case it is easy to verify that the structures listed are homogeneous, but hard to show that the list is complete.

We can represent L_0 -structures by diagrams in which $a \circ \rightarrow \circ b$ means “ $(a, b) \in R$ and $(b, a) \notin R$,” and $a \circ \longleftrightarrow b$ means “ $(a, b), (b, a) \in R$.” Let \mathcal{A}, \mathcal{B} , and \mathcal{C} denote $\circ \rightarrow \circ$, $\circ \rightarrow \circ \rightarrow \circ$, and the 3-cycle respectively. Let \mathcal{L} denote the loop .

Let $\mathbf{G} \subseteq \mathbf{S}(L_0)$ be the strictly universal class constrained by the structures \mathcal{A} and \mathcal{L} . Then \mathbf{G}^c is the class of countable *graphs*. We will list the structures in $\mathbf{G}^c \cap \text{Hom}(L)$, which is the same thing as listing the amalgamation classes contained in \mathbf{G} .

Let K_n denote the complete graph on n vertices and I_G the diagonal of $G \times G$. For any graph $\mathcal{G} = (G, R^{\mathcal{G}})$, write $\bar{\mathcal{G}}$ for its *complement*, namely,

$$(G, (G \times G) \setminus (R^{\mathcal{G}} \cup I_G)).$$

The complement of K_n is denoted I_n . For any graphs $\mathcal{G}_0, \mathcal{G}_1$, let $\mathcal{G}_0[\mathcal{G}_1]$ be the wreath product obtained by replacing each vertex of \mathcal{G}_0 by a copy of \mathcal{G}_1 , let $\mathcal{G}_0 \times \mathcal{G}_1$ be the usual cartesian product, and let $\mathcal{G}_0 + \mathcal{G}_1$ be the disjoint union. Let

$$\mathcal{P}e = (\{0, 1, 2, 3, 4\}, \{(i, j) : |i - j| \in \{1, 4\}\})$$

be the pentagon.

TABLE 1. Homogeneous graphs

Graph \mathcal{M}	Constraints of $\mathbf{S}(\mathcal{M})$
$\mathcal{P}e$	$K_3, I_3, K_2 \times K_2, \overline{K_2 \times K_2}$
$K_3 \times K_3$	$K_4, I_4, K_1 + K_3, \overline{K_1 + K_3}, K_2 + I_2, \overline{K_2 + I_2}$
$I_m[K_n]$	$I_{m+1}, K_{n+1}, \overline{K_1 + K_2}$
$I_\omega[K_n]$	$K_{n+1}, \overline{K_1 + K_2}$
$I_m[K_\omega]$	$I_{m+1}, \overline{K_1 + K_2}$
$I_\omega[K_\omega]$	$\overline{K_1 + K_2}$
$\mathcal{G}(m)$	K_{m+1}
$\mathcal{G}(\omega)$	none

Table 1 lists all homogeneous graphs up to complements; m, n run through the positive integers. The sets of constraints are relative to \mathbf{G} , i.e., to each set should be added the constraints of \mathbf{G} . We have not given explicit constructions of the “generic” graphs $\mathcal{G}(m)$ and $\mathcal{G}(\omega)$. However, from Theorem 1, specifying the amalgamation classes fixes these graphs up to isomorphism. The sources for this table are [G, W2, LW].

Let $\mathbf{D} \subseteq \mathbf{S}(L_0)$ be the universal class constrained by \mathcal{L} and K_2 . Then \mathbf{D}^c is the class of countable *directed graphs*. For each $M \in \mathbf{D}$, $\#M$ is obtained by adjoining a new vertex which dominates each vertex of M , while $M^\#$ is obtained by adjoining a new vertex dominated by each vertex of M . Within \mathbf{D} , let \mathbf{T} be the universal class constrained by I_2 , and \mathbf{P} the universal class constrained by \mathcal{B} and \mathcal{C} . Then \mathbf{T}^c and \mathbf{P}^c are the classes of countable tournaments and partial orders respectively. Let the partial order of the rational numbers be denoted \mathcal{Q} .

Schmerl [S] established the list of homogeneous partial orders presented in Table 2.

TABLE 2. Homogeneous partial orders

Partial order M	Constraints of $\mathbf{S}(M)$
I_m	\mathcal{A}, I_{m+1}
$I_m[\mathcal{Q}]$	$\#I_2, I_2^\#, I_{m+1}$
$\mathcal{Q}[I_m]$	$\mathcal{A} + I_1, I_{m+1}$
$I_\omega[\mathcal{Q}]$	$\#I_2, I_2^\#$
$\mathcal{Q}[I_\omega]$	$\mathcal{A} + I_I$
\mathcal{P}	none

As before $1 \leq m < \omega$ and the constraints of \mathbf{P} have been omitted. The wreath products are defined in the same way as for graphs. \mathcal{P} is the “generic” partial order. The classifications of homogeneous partial orders by Schmerl, and of homogeneous graphs by Woodrow and the author, confirmed conjectures of Henson [H2]. The most important contribution of Henson’s paper was showing there are 2^{\aleph_0} homogeneous directed graphs; the same was shown by Peretyatkin [P] independently. We will return to this point below.

The author [L3] determined all homogeneous tournaments. They are listed in Table 3.


TABLE 3. Homogeneous tournaments

Tournament M	Constraints of $\mathbf{S}(M)$
I_1	\mathcal{A}
\mathcal{C}	$\mathcal{A}^\#$
\mathcal{Q}	\mathcal{C}
\mathcal{Q}^*	$\mathcal{C}^\#, \# \mathcal{C}$
\mathcal{T}	none

The most interesting of these structures is \mathcal{Q}^* , called a “dense local order” by Cameron [Ca, §6]. The first reference to \mathcal{Q}^* we know of is [W1, p. 53]. We now come to the recent work of Cherlin [C1, C2], which characterizes all

homogeneous directed graphs except possibly for some primitive ones embedding I_ω . Cherlin conjectures that there are at most countably many homogeneous directed graphs not in his catalogue. The following tables summarize Cherlin’s work, which subsumes the earlier work on partial orders and tournaments. A directed graph M is *deficient* if at least one of \mathcal{A} and I_2 is not embeddable in M . In Table 4 are listed all the deficient and imprimitive homogeneous directed graphs. Table 5 lists all other *known* homogeneous directed graphs.

TABLE 4. Deficient and imprimitive homogeneous directed graphs

Directed graph M	Constraints of $S(M)$
$I_m[\mathcal{X}]$	$\# I_2, I_2^\#, \mathcal{B}, I_{m+1}$, constraints of \mathcal{X}
$I_\omega[\mathcal{X}]$	$\# I_2, I_2^\#, \mathcal{B}$, constraints of \mathcal{X}
$\mathcal{X}[I_m]$	$\mathcal{A} + I_1, \mathcal{B}, I_{m+1}$, constraints of \mathcal{X}
$\mathcal{X}[I_\omega]$	$\mathcal{A} + I_1, \mathcal{B}$, constraints of \mathcal{X}
\mathcal{Y}^\wedge	$I_3, \mathcal{A} + I_1, \# Z, Z^\#$ (Z constraint of \mathcal{Y})
$m * I_\omega$	$\mathcal{A} + I_1$, all tournaments of size $m + 1$
$\omega * I_\omega$	$\mathcal{A} + I_1$
S	$\mathcal{A} + I_1$, 

Key : \mathcal{X}, \mathcal{Y} are homogeneous tournaments, $\mathcal{Y} \neq \mathcal{Q}^* : 1 \leq m < \omega$.

The interesting entries in Table 4 are the \mathcal{Y}^\wedge and the “semigeneric” directed graph S . I_1^\wedge and C^\wedge had turned up previously in [L1].

TABLE 5. Known nondeficient primitive homogeneous directed graphs

Directed graph M	Constraints of $S(M)$
\mathcal{Q}^0	$C, I_2^\#, \# I_2$
$\mathcal{D}(m)$	I_{m+1}
$\mathcal{D}(\mathbf{A})$	Z ($Z \in \mathbf{A}$)

Key : \mathbf{A} is any antichain in \mathbf{T} ; $2 \leq m < \omega$.

As mentioned above Table 5 is only known to be complete with respect to directed graphs not embedding I_ω . The surprising entry here is \mathcal{Q}^0 , which Cherlin has called the “myopic local order.” The final entry in Table 5 can be seen as stemming from the discussion in [H2]. Henson showed that there is an infinite antichain $\mathbf{A} \subseteq \mathbf{T}$. From this he deduced that there are 2^{\aleph_0} distinct amalgamation classes $\subseteq \mathbf{D}$ because every subset \mathbf{B} of \mathbf{A} constrains a different one. Hence there are 2^{\aleph_0} pairwise nonisomorphic homogeneous directed graphs.

This concludes our account of the progress that has been made in classifying the homogeneous L_0 -structures. Lack of space precludes us from attempting to describe how these results are proved.

3. Stable homogeneous structures. An L -structure M is *unstable* if for some N elementarily equivalent to M there exist $n < \omega$, n -tuples $\bar{a}_i \in N$ ($i < \omega$), and an L -formula $\varphi(\bar{x}, \bar{y})$ such that $N \models \varphi(\bar{a}_i, \bar{a}_j)$ iff $i < j$. Otherwise, M is *stable*.

In [CL, Theorem 1] the following theorem is proved by applying the theory of permutation groups. The same result for binary languages was proved earlier in [SL] using purely model-theoretic methods.

THEOREM 2. *Let L be a finite relational language. There is an L -sentence σ such that for all $M \in \text{Hom}(L)$, M is stable if and only if $M \models \sigma$.*

From [CHL, Corollary 7.4] this theorem is easily seen to be equivalent to:

THEOREM 2'. *Let L be a finite relational language. For all $M \in \text{Hom}(L)$, M is stable if and only if M is the union of a chain of finite homogeneous structures.*

Shrinking. Consider $M \in \text{Hom}(L)$. Let $\text{Th}(N)$ be \aleph_0 -categorical and $M \subseteq N$. Then N is an *extension by definitions* of M if M is 0-definable in N , $\{b\}$ is M -definable in N for each $b \in N$, and $\text{Aut}(M) = \{\alpha \mid M: \alpha \in \text{Aut}(N)\}$. Suppose that in some extension by definitions of M there is an invariant family Ψ of pairwise disjoint, definable, infinite indiscernible sets. Let $\text{Aut}(M)$ act transitively on Ψ . Suppose further that there is an invariant mapping C of M into the finite subsets of $\bigcup \Psi$ such that for all $b_0, b_1 \in I \in \Psi$ there exists $\alpha \in \text{Aut}(M)$ such that $\alpha(b_0) = b_1, \alpha(b_1) = b_0$, and

$$((\bigcup \Psi) \setminus \{b_0, b_1\}) \cup \{a \in M: b_0, b_1 \notin C(a)\} \subseteq \text{Fix}(\alpha).$$

In this case we call Ψ a *nice family* attached to M . Two nice families Ψ_0, Ψ_1 are *equivalent* if there is an invariant bijection between $\bigcup \Psi_0$ and $\bigcup \Psi_1$. The number of inequivalent nice families attached to M is bounded in terms of L .

If Ψ is a nice family attached to M , we can *shrink M with respect to Ψ* as follows. Choose $m < \omega$, the *target dimension*, and $B \subseteq \bigcup \Psi$ such that $|B \cap I| = m$ for all $I \in \Psi$. Let $N \subseteq M$ be the substructure with universe $\{a \in M: C(a) \subseteq B\}$. Then N is said to be obtained by shrinking M . $N \in \text{Hom}(L)$ and is fixed by m up to an automorphism of M . There is no difficulty in shrinking M simultaneously with respect to several inequivalent nice families, each with its own target dimension. More details about shrinking can be found in [L2, §12; L4, and KL].

EXAMPLE. $M = I_\omega[K_\omega]$. Let E be the equivalence relation on M whose classes are the copies of K_ω . There are two nice families: $\Psi_0 = M/E$ and $\Psi_1 = \{M/E\}$. Shrinking M with respect to Ψ_0 gives $I_\omega[K_m]$, and with respect to Ψ_1 gives $I_m[K_\omega]$.

One of the main results of [L2], Theorem 15.3, says that Theorem 2 above implies:

THEOREM 3. *Let L be a finite relational language. There exists a finite subclass $\mathbf{H} \subseteq \text{Hom}(L)$ such that, for all $M \in \text{Hom}(L)$, M is stable iff there exists $N \cong M$ such that either $N \in \mathbf{H}$ or N is obtained by shrinking a member of \mathbf{H} .*

For the next result we need an additional piece of notation. For any L -structure M and $k < \omega$ let $\mathbf{U}(M, k)$ denote the class of all finite L -structures \mathcal{F} such that every substructure of \mathcal{F} of size $\leq k$ is embeddable in M . The following theorem is due to Harrington and can be used to give a simpler proof of Theorem 3 than is afforded by [L2].

THEOREM 4. *Let L be a finite relational language. There exists n such that for all stable $M \in \text{Hom}(L)$, $\mathbf{U}(M, k)$ is an amalgamation class for some $k < n$.*

A proof of Harrington's theorem together with the resulting simplified proof of Theorem 2 will appear in [KL]. Combining this theorem with the theory of shrinking and dimensions developed in [L2, §§11 and 12] we immediately obtain the following theorem, which may be thought of as Theorem 4 in another guise. We call an amalgamation class *stable* if the structure associated with it by Theorem 1 is stable.

THEOREM 5. *A stable amalgamation class over a finite relational language is finitely constrained.*

4. Some conjectures. Fix a finite relational language L for this section.

Let $\mathbf{A}, \mathbf{B} \subseteq \mathbf{F}(L)$ be finite. We write

$$\bigwedge \mathbf{A} \Rightarrow \bigvee \mathbf{B}$$

if every amalgamation class over L which includes \mathbf{A} intersects \mathbf{B} . Since $\text{Th}(\text{Hom}(L))$ is axiomatizable, the relation \Rightarrow is recursively enumerable. One goal of the theory of homogeneous structures is to prove:

CONJECTURE 1. \Rightarrow is recursive.

Note that the theory of stable homogeneous structures sketched ever so lightly in §3 shows that the corresponding relation obtained by restricting to finite amalgamation classes is recursive.

The following is immediate:

LEMMA. *The union and intersection, when nonempty, of a chain of amalgamation classes over L are also amalgamation classes.*

The single most important question about $\text{Hom}(L)$ is addressed by:

CONJECTURE 2. Every amalgamation class over L is the intersection of a chain of finitely constrained amalgamation classes.

If L is binary Conjecture 2 implies Conjecture 1. Indeed, the algorithm Conjecture 2 suggests for deciding " $\bigwedge \mathbf{A} \Rightarrow \bigvee \mathbf{B}$?" is valid, even if Conjecture 2 fails, provided every pair (\mathbf{A}, \mathbf{B}) which can be separated by an amalgamation class can be separated by a finitely constrained amalgamation class.

Let $U \subseteq F(L)$ be a strictly universal class. If $U \neq \emptyset$, there is certainly a finitely constrained amalgamation class $A \subseteq U$ corresponding to a one-point structure. Our final piece of speculation is:

CONJECTURE 3. If $U \subseteq F(L)$ is a nonempty strictly universal class, then the number of maximal finitely constrained amalgamation classes $\subseteq U$ is finite.

REFERENCES

- [A] C. Ash, *Undecidable \aleph_0 -categorical theories*, Notices Amer. Math. Soc. **17** (1970), 295, #70T-E10.
- [Ca] P. J. Cameron, *Orbits of permutation groups on unordered sets. II*, J. London Math. Soc. (2) **20** (1981), 249–264.
- [C1] G. L. Cherlin, *Homogeneous directed graphs: the imprimitive case*, Logic Colloquium '85 (to appear).
- [C2] —, *Homogeneous directed graphs omitting I_∞* (in preparation).
- [CHL] G. Cherlin, L. Harrington, and A. H. Lachlan, *\aleph_0 -categorical, \aleph_0 -stable structures*, Ann. Pure Appl. Logic **28** (1985), 103–135.
- [CL] G. Cherlin and A. H. Lachlan, *Stable finitely homogeneous structures*, Trans. Amer. Math. Soc. **296** (1986), 815–850.
- [F] R. Fraïssé, *Sur l'extension aux relations de quelques propriétés des ordres*, Ann. Sci. École Norm. Sup. **71** (1954), 361–388.
- [G] A. Gardiner, *Homogeneous graphs*, J. Combin. Theory Ser. B **20** (1976), 94–102.
- [G1] W. Glassmire, *A problem in categoricity*, Notices Amer. Math. Soc. **17** (1970), 295, #70T-E7.
- [H1] C. W. Henson, *A family of countable homogeneous graphs*, Pacific J. Math. **38** (1971), 69–83.
- [H2] —, *Countable homogeneous relational structures and \aleph_0 -categorical theories*, J. Symbolic Logic **37** (1972), 494–500.
- [KL] J. Knight and A. H. Lachlan, *Shrinking, stretching, and codes for homogeneous structures* (in preparation).
- [L1] A. H. Lachlan, *Finite homogeneous simple digraphs*, Proceedings of the Herbrand Symposium (Marseilles, 1981), J. Stern, editor, Stud. Logic Foundations Math., vol. 107, North-Holland, 1982, pp. 89–108.
- [L2] —, *On countable stable structures which are homogeneous for a finite relational language*, Israel J. Math. **49** (1984), 69–153.
- [L3] —, *Countable homogeneous tournaments*, Trans. Amer. Math. Soc. **284** (1984), 431–461.
- [L4] —, *Binary homogeneous structures. I*, Proc. London Math. Soc. (3) **52** (1986), 385–411.
- [L5] —, *Binary homogeneous structures. II*, Proc. London Math. Soc. (3) **52** (1986), 412–426.
- [LS] A. H. Lachlan and S. Shelah, *Stable structures homogeneous for a finite binary language*, Israel J. Math. **49** (1984), 155–180.
- [LW] A. H. Lachlan and R. E. Woodrow, *Countable ultrahomogeneous graphs*, Trans. Amer. Math. Soc. **262** (1980), 51–94.
- [P] M. G. Peretyatkin, *On complete theories with a finite number of denumerable models*, Algebra and Logic **12** (1973), 310–326.
- [S] J. Schmerl, *Countable homogeneous partially ordered sets*, Algebra Universalis **9** (1979), 317–321.
- [W1] R. E. Woodrow, *Theories with a finite number of countable models and a small language*, Ph. D. Thesis, Simon Fraser University, 1976.
- [W2] —, *There are four countable ultrahomogeneous graphs without triangles*, J. Combin. Theory Ser. B **27** (1979), 168–179.

Конечно аксиоматизируемые теории

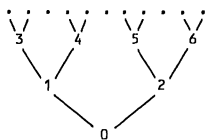
М. Г. ПЕРЕТЯТКИН

В работах автора [7] и [8] описаны две конструкции, позволяющие строить конечно аксиоматизируемые теории с самыми различными свойствами алгоритмического характера. Кроме теорем существования в этих работах получены оценки сложности различных естественных классов предложений. Вторая конструкция, использующая в качестве базисной теорию работы [6], имеет значительно большую область применимости, зато первая сравнительно более простая. В данной работе приводится ряд результатов о конечно аксиоматизируемых теориях, расширяющих и дополняющих содержание работ [7, 8].

Предварительные сведения. Все теории рассматриваются в логике предикатов первого порядка с равенством. Общие понятия, используемые в работе, соответствуют монографиям Кейслера-Чэна [4], Роджерса [9], Ершова [3]. Язык будем называть богатым, если он содержит хотя бы один предикатный символ арности два или больше. Кроме специально оговоренных случаев, все результаты предполагаются относящимися к произвольному конечному богатому языку.

Нумерацией модели \mathcal{M} называется отображение $\nu: N \xrightarrow{\text{на}} |\mathcal{M}|$. Модель \mathcal{M} с нумерацией ν называется нумерованной моделью и обозначается (\mathcal{M}, ν) . Нумерованная модель (\mathcal{M}, ν) называется конструктивной, если все её предикаты и функции рекурсивны в нумерации ν , и называется сильно конструктивной, если $\text{Th}(\mathcal{M}_\nu)$ разрешима, где \mathcal{M}_ν — обогащение языка модели \mathcal{M} счётным множеством новых констант c_i , $i < \omega$, причём c_i интерпретируется элементом $\nu(i)$. Модель \mathcal{M} называется автоустойчивой, если для любых её сильных конструктивизаций ν_1 и ν_2 существует рекурсивный изоморфизм $\mu: (\mathcal{M}, \nu_1) \rightarrow (\mathcal{M}, \nu_2)$. Нуртазин [5] показал, что если полная разрешимая теория T имеет сильно конструктивизируемую простую модель \mathcal{M} , то \mathcal{M} будет автоустойчивой тогда и только тогда, когда множество атомов алгебр Линденбаума теории T рекурсивно.

Полным бинарным деревом назовём частично упорядоченное множество $\mathcal{D}_0 = \langle N; \preceq \rangle$ формы, изображённой на рисунке. Имеются две



естественные операции взятия левого и правого последователя, которые задаются формулами: $L(x) = 2x + 1$, $R(x) = 2x + 2$. Деревом назовём всякое множество $\mathcal{D} \subseteq N$, удовлетворяющее следующим двум условиям:

1. $m \leq n$, $n \in \mathcal{D} \rightarrow m \in \mathcal{D}$,
2. $L(n) \in \mathcal{D} \leftrightarrow R(n) \in \mathcal{D}$ для всех $n \in N$.

Естественно определены понятия тупика дерева, цепи, максимальной цепи дерева, изолированной цепи в заданном семействе цепей. Через $\Pi(\mathcal{D})$ обозначается семейство всех максимальных цепей дерева \mathcal{D} , а через $\Pi^{\text{fin}}(\mathcal{D})$ — семейство всех конечных максимальных цепей дерева \mathcal{D} . Дерево называется атомным, если над каждым его элементом есть хотя бы один тупик. Дерево называется суператомным, если $\Pi(\mathcal{D})$ не более, чем счётное. Через $\Pi_\alpha(\mathcal{D})$ обозначим α -кратную итерацию выбрасывания изолированных цепей, начиная с $\Pi(\mathcal{D})$. Отсюда естественно получается определение ранга цепи и ранга дерева. Можно доказать, что дерево будет суператомным тогда и только тогда, когда $\Pi_\alpha(\mathcal{D}) = \emptyset$ при некотором α , т.е. когда все цепи в $\Pi(\mathcal{D})$ имеют ранги. Более подробное изложение всех перечисленных выше понятий, связанных с деревьями, имеется в [8].

W_s , $s < \omega$ — постовская нумерация всех р.п. множеств, W_s^A , $s < \omega$ — нумерация множеств, рекурсивно перечислимых относительно множества $A \subseteq N$. Через $[B]_{\mathcal{D}}$ обозначим замыкание множества $B \subseteq N$ до дерева. Будем использовать обозначения

$$\mathcal{D}_s = [W_s]_{\mathcal{D}}, \quad s < \omega, \quad \mathcal{D}_s^A = [W_s^A]_{\mathcal{D}}, \quad s < \omega, \quad A \subseteq N.$$

Через ε_k , $k < \omega$, обозначим табличное условие утверждающее, что «множество содержит элемент k ». Табличным условием (*tt*-условием) будем называть пропозициональную формулу, построенную из элементарных высказываний ε_k . Утверждение, что табличное условие истинно на множестве A будем символически записывать в виде $A \models \tau$. Посредством τ_k , $k < \omega$, будем обозначать фиксированную гёделевскую нумерацию всех табличных условий. Ещё одно специальное определение

$$\mathfrak{A}_m = \{A \subseteq N \mid (\forall k \in W_m) A \models \tau_k\}, \quad m < \omega.$$

Следующая общая теорема является усилением теоремы 9.1 работы [8] посредством дополнительного пункта об автоустойчивости простой модели.

ТЕОРЕМА 1 (ОСНОВНАЯ ТЕОРЕМА). Пусть \mathcal{L} — произвольный конечный богатый язык. Эффективно по \mathcal{L} и по заданной паре натуральных чисел $\langle m, s \rangle$ строится конечно аксиоматизируемая модельно

полная теория $F(m, s)$ языка \mathcal{L} и рекурсивная последовательность Ψ_n , $n \in N$, предложений языка \mathcal{L} , имеющие свойства:

1. Предложения Ψ_n , $n \in N$, порождают алгебру Линденбаума теории $F(m, s)$.

2. Теория

$$F(m, s)[A] = F(m, s) \cup \{\Psi_i \mid i \in A\} \cup \{\neg\Psi_j \mid j \in N \setminus A\}, \quad A \subseteq N,$$

является непротиворечивой тогда и только тогда, когда $A \in \mathfrak{R}_m$.

3. При любом $A \in \mathfrak{R}_m$ справедливы соотношения

(а) Теория $F(m, s)[A]$ имеет простую модель \leftrightarrow дерево \mathcal{D}_s^A является атомным.

(б) Простая модель теории $F(m, s)[A]$, если она существует, является сильно конструктивизируемой \leftrightarrow множество A рекурсивно и семейство $\Pi^{\text{fn}}(\mathcal{D}_s^A)$ вычислимо.

(в) Простая модель теории $F(m, s)[A]$, если она существует и сильно конструктивизируема, является автоустойчивой относительно сильных конструктивизаций \leftrightarrow дерево \mathcal{D}_s^A является рекурсивным.

(г) Теория $F(m, s)[A]$ имеет счётную насыщенную модель \leftrightarrow дерево \mathcal{D}_s^A является суператомным.

(д) Счётная насыщенная модель теории $F(m, s)[A]$ сильно конструктивизируема \leftrightarrow множество A рекурсивно и семейство $\Pi(\mathcal{D}_s^A)$ вычислимо.

(е) Теория $F(m, s)[A]$ является ω — стабильной \leftrightarrow дерево \mathcal{D}_s^A суператомное.

(ж) Ранг Морли теории $F(m, s)[A]$ равен

$$\max\{33, 1 + \text{Rank } \mathcal{D}_s^A + \gamma\},$$

где $\gamma = 2$, если дерево \mathcal{D}_s^A является суператомным и $\gamma = 0$, если дерево \mathcal{D}_s^A — не суператомное.

ДОКАЗАТЕЛЬСТВО. Придерживаясь системы обозначений работы [8], докажем (в).

Пусть A и \mathcal{D}_s^A — рекурсивные, (\mathfrak{M}, ν) , (\mathfrak{M}_1, ν_1) — две сильно конструктивных простых модели теории $F(m, s)[A]$. Существование эффективного соответствия между каркасами этих моделей очевидно. Поэтому всё сводится к нахождению рекурсивного изоморфизма на множестве W . Так как модель простая, то любой элемент $a \in W \cap U$ порождает компоненту, которая определяет K -путь, соответствующий некоторому тупику дерева \mathcal{D}_s^A . Переход от a к указанному тупику дерева \mathcal{D}_s^A , очевидно, эффективный. Это позволяет построить рекурсивный изоморфизм данных нумерованных моделей.

Обратно, если A — рекурсивное, но \mathcal{D}_s^A — не рекурсивное, то множество атомных формул теории $F(m, s)[A]$ с одной свободной переменной не может быть рекурсивным. Тогда по критерию работы [5] простая модель этой теории не автоустойчива относительно сильных конструктивизаций. \square

Из теоремы 1 легко выводится ещё одна общая теорема, касающаяся полных теорий:

ТЕОРЕМА 2 (\mathcal{D} -ВАРИАНТ ОСНОВНОЙ ТЕОРЕМЫ). Пусть \mathcal{L} — конечный богатый язык, \mathcal{D} — рекурсивно перечислимое дерево. Эффективно по L и по р.п. индексу дерева \mathcal{D} может быть построена полная, модельно полная конечно аксиоматизируемая теория $F(\mathcal{D})$ языка \mathcal{L} , имеющая следующие свойства:

- (а) Теория $F(\mathcal{D})$ имеет простую модель \leftrightarrow дерево \mathcal{D} является атомным.
- (б) Простая модель теории $F(\mathcal{D})$, если она существует, является сильно конструктивизируемой \leftrightarrow семейство $\Pi^{\text{fn}}(\mathcal{D})$ вычислимо.
- (в) Простая модель теории $F(\mathcal{D})$, если она существует и сильно конструктивизируема, является автоустойчивой относительно сильных конструктивизаций \leftrightarrow дерево \mathcal{D} является рекурсивным.
- (г) Теория $F(\mathcal{D})$ имеет счётную насыщенную модель \leftrightarrow дерево \mathcal{D} является суператомным.
- (д) Счётная насыщенная модель теории $F(\mathcal{D})$ сильно конструктивизируема \leftrightarrow семейство $\Pi(\mathcal{D})$ вычислимо.
- (е) Теория $F(\mathcal{D})$ является ω -стабильной \leftrightarrow дерево \mathcal{D} суператомное.
- (ж) Ранг Морли теории $F(\mathcal{D})$ равен

$$\max\{33, 1 + \text{Rank } \mathcal{D}_s^A + \gamma\},$$

где $\gamma = 2$, если дерево \mathcal{D} является суператомным, и $\gamma = 0$, если дерево \mathcal{D} — не суператомное.

ДОКАЗАТЕЛЬСТВО. Выберем и зафиксируем m так, чтобы $\mathfrak{A}_m = \emptyset$. В качестве искомой теории $F(\mathcal{D})$ можно взять теорию $F(m, s)$, где s выбрано таким образом, что $\mathcal{D}_s^\emptyset = \mathcal{D}$. \square

Следующая теорема, характеризующая алгебры Линденбаума конечно аксиоматизируемых теорий, равноценна доказанным в [7] и [8] утверждениям, решающим известную проблему Ханфа работы [11]. Такой же результат анонсирован Ханфом [12], однако доказательство им не опубликовано.

ТЕОРЕМА 3. Пусть \mathcal{L} — произвольный конечный богатый язык. Нумерованная булева алгебра (\mathcal{B}, ν) рекурсивно эквивалентна алгебре Линденбаума некоторой конечно аксиоматизируемой теории F языка \mathcal{L} тогда и только тогда, когда (\mathcal{B}, ν) является позитивно нумерованной алгеброй.

ДОКАЗАТЕЛЬСТВО. Если (\mathcal{B}, ν) рекурсивно эквивалентна алгебре Линденбаума некоторой конечно аксиоматизируемой теории, то (\mathcal{B}, ν) будет, очевидно, позитивно нумерованной алгеброй. Поэтому достаточно показать, как строить F по заданной позитивно нумерованной булевой алгебре (\mathcal{B}, ν) .

Обозначим через \mathfrak{B} булеву алгебру, элементами которой являются классы эквивалентных табличных условий, а операции индуцированы

пропозициональными связками. Очевидно, что \mathfrak{B} является безатомной алгеброй, свободно порождённой элементарными табличными условиями. Алгебра \mathfrak{B} имеет естественную конструктивизацию μ , определяемую гёделевской нумерацией табличных условий. Из пунктов 1, 2 основной теоремы следует, что алгебра Линденбаума теории $F(m, s)$ рекурсивно эквивалентна фактор-алгебре

$$(\mathfrak{B}/\mathcal{F}_m, \mu^*) \quad (1)$$

где \mathcal{F}_m — фильтр, порождённый множеством $\{\tau_k \mid k \in W_m\}$, а μ^* — нумерация фактор-алгебры, индуцированная нумерацией μ . Нетрудно видеть, что каждый рекурсивно перечислимый фильтр алгебры \mathfrak{B} совпадает с \mathcal{F}_m для некоторого $m < \omega$. Отсюда получаем нужное утверждение, т.к. в форме (1) представима с точностью до изоморфизма всякая позитивно нумерованная булева алгебра. \square

Следующая теорема даёт некоторую конкретную информацию о возможностях алгебр Линденбаума конечно аксиоматизируемых теорий.

ТЕОРЕМА 4. Пусть \mathcal{L} — произвольный конечный богатый язык, тогда

(а) Алгебра Линденбаума теории, определяемой пустым множеством аксиом в языке \mathcal{L} , является неконструктивизируемой.

(б) Существует разрешимая конечно аксиоматизируемая теория языка \mathcal{L} с конструктивизируемой, но не сильно конструктивизируемой алгеброй Линденбаума любого из элементарных типов, кроме категоричных в мощностях $\delta \leq \omega$.

ДОКАЗАТЕЛЬСТВО. (а) Обозначим названную алгебру через $\mathfrak{B}(\mathcal{L})$. В работе [10] построена позитивно нумерованная булева алгебра, которая не является конструктивизируемой. По теореме 3 существует конечно аксиоматизируемая теория F заданного языка, имеющая неконструктивизируемую алгебру Линденбаума. Нетрудно видеть, что алгебра Линденбаума $\mathfrak{B}(F)$ изоморфна фактор-алгебре $\mathfrak{B}(\mathcal{L})$ по некоторому главному фильтру. Отсюда вытекает неконструктивизируемость алгебры $\mathfrak{B}(\mathcal{L})$.

(б) В работе [1] построена конструктивизируемая, но не сильно конструктивизируемая булева алгебра любого из элементарных типов, указанных в теореме. Искомая конечно аксиоматизируемая теория получается по теореме 3. \square

Следующая теорема решает две проблемы Харрингтона работы [13], а также проблему работы [7].

ТЕОРЕМА 5. (а) Существует полная конечно аксиоматизируемая ω -стабильная теория конечного ранга Морли, у которой простая модель и счётная насыщенная модель не являются конструктивизируемыми.

(б) Существует полная конечно аксиоматизируемая ω -стабильная теория конечного ранга Морли, простая модель которой сильно конструктивизируема, но не автоустойчива относительно сильных конструктивизаций.

ДОКАЗАТЕЛЬСТВО. (а) Возьмём построенное в работе [2] рекурсивно перечислимое суператомное дерево \mathcal{D} , для которого семейства $\Pi(\mathcal{D})$ и $\Pi^{\text{fn}}(\mathcal{D})$ не вычислимы. По построению ранг этого дерева — конечный. Из модельной полноты теории $F(\mathcal{D})$ следует, что конструктивизируемость её моделей равносильна сильной конструктивизируемости. Поэтому теория $F(\mathcal{D})$ будет по теореме 2 искомой.

(б) Указанные свойства имеет теория $F(\mathcal{D}')$, где \mathcal{D}' — рекурсивно перечислимое нерекурсивное суператомное дерево конечного ранга, для которого семейство $\Pi^{\text{fn}}(\mathcal{D}')$ является вычислимым. \square

Следующая теорема, анонсированная в [8], усиливает результат Лахлана [14], доказанный для счётных теорий. Фактически построение Лахлана даёт разрешимую теорию.

ТЕОРЕМА 6. *Существует полная конечно аксиоматизируемая теория T ранга Морли $\alpha_T = \omega_1$.*

ДОКАЗАТЕЛЬСТВО. Здесь приводится схематичное доказательство, для понимания которого необходимо детальное знакомство с конструкциями работ [7] и [8]. Для построения искомой теории T используется новая конструкция, при характеристизации которой мы будем использовать терминологию работы [8].

Каркас теории T имеет форму сетки, построенной на основе теории квазиследования [6]. В теории функционирует бесконечное множество машин Тьюринга, использующих информацию одного общего оракула. При этом, информация оракула должна быть размещена без промежутков, в последовательных ячейках бесконечной в одну сторону ленты, ограниченной справа точкой O . Вправо от этой точки располагается рабочая область.

Одна из машин является активной по отношению к оракулу. Она должна формировать его содержимое таким образом, чтобы на конечных удалениях влево от точки O встречались всевозможные конечные комбинации последовательно расположенных нулей и единиц. В результате, по теореме компактности, в нестандартных частях оракула могут содержаться произвольные последовательности нулей и единиц.

Все пассивные машины имеют одну и ту же программу, но конструкция каркаса теории T должна обеспечивать, чтобы все они работали в разных плоскостях и воспринимали содержимое оракула с различными сдвигами относительно точки O . Каждая пассивная машина выполняет работу аналогичную той, которая описана в конструкции работы [8]. Грубо говоря, программа пассивной машины должна обеспечивать переработку, путём деления клеток ленты, содержимого оракула A в дерево $[A]_{\mathcal{D}}$, которое должно быть связано с транслятором, влияющим на ранг Морли соответствующих формул. Так как содержимым оракула пассивной машины может оказаться произвольное множество $A \subseteq N$, то соответствующее дерево $[A]_{\mathcal{D}}$ может быть суператомным любого

счётного ранга. В результате в теории T будут существовать формулы с параметрами сколь угодно большого счётного ранга. Например, такими будут формулы с параметрами, взятыми из нестандартных частей оракула. Таким образом, мы получаем $\alpha_T = \omega_1$. \square

Теперь, используя эффективность построения конечно аксиоматизируемых теорий, гарантированную основной теоремой, исследуем сложность двух важных классов предложений. Первым рассмотрим класс предложений, изучавшийся в работе Вота [15], где для него была получена нижняя оценка Π_1^0 .

ТЕОРЕМА 7. Пусть FA — класс моделей заданного конечного богатого языка, имеющих конечно аксиоматизируемую теорию. Тогда $Th(FA) \approx \Pi_3^0$.

ДОКАЗАТЕЛЬСТВО. Множество $Th(FA)$ рекурсивно изоморфно дополнению множества

$$L = \{n \mid \Phi_n \text{ имеет к.а. пополнение}\}.$$

Поэтому достаточно показать, что $L \approx \Sigma_3^0$. Имеем

$$n \in L \leftrightarrow (\exists \Psi)(\Phi_n \& \Psi \text{ — полная теория}).$$

Учитывая, что свойство «полная теория» описывается префиксом формы $\forall \exists$, получаем для L верхнюю оценку Σ_3^0 . Для нижней оценки возьмём эталонное Σ_3^0 -множество

$$S = \{n \mid N \setminus W_n \text{ — конечное}\}$$

[9, §14.8]. Теория T_n языка $\mathcal{L} = \{P_0^1, P_1^1, \dots, P_s^1, \dots\}$, определяемая следующим рекурсивным множеством аксиом

$$\{(\exists x)(\forall y)x = y\} \cup \{(\forall x)P_i(x), i \in W_m\},$$

будет иметь конечно аксиоматизируемое пополнение тогда и только тогда, когда $n \in S$. Пусть $f(x)$ — о.р.ф. такая, что предложение $\Phi_{f(n)}$ определяет конечно аксиоматизируемую теорию с алгеброй Линденбаума, изоморфной алгебре Линденбаума теории T_n . Тогда выполняется

$$n \in S \leftrightarrow \Phi_{f(n)} \text{ имеет к.а. пополнение.}$$

Это даёт необходимую нижнюю оценку для L .

Ещё одна оценка сложности класса формул.

ТЕОРЕМА 8. $\{n \mid \Phi_n \text{ определяет полную стабильную теорию}\} \approx \Pi_2^0$.

ДОКАЗАТЕЛЬСТВО. Используя известный результат Шелаха о характеристизации стабильных теорий [4, 7.1.33], получаем, что n принадлежит указанному в теореме множеству тогда и только тогда, когда Φ_n определяет полную теорию, причём в этой теории нет свойства линейного порядка. Каждое из этих условий описывается префиксом формы $\forall \exists$. Для нижней оценки возьмём эталонное Π_2^0 -множество, определённое в [9, §14.8],

$$I = \{n \mid W_n \text{ — бесконечно}\}.$$

Рассмотрим язык \mathcal{L}' содержащий только предикат равенства. Через Δ_k обозначим предложение, утверждающее, что существует не менее k

различных элементов. Рассмотрим теорию T_n языка \mathcal{L}' , определяемую множеством аксиом $\{\Delta_k \mid k \in W_n\}$. Легко видеть, что эта теория полна тогда и только тогда, когда $n \in I$. Эффективно по n построим конечно аксиоматизируемую теорию F такую, что $\mathfrak{B}(T_n) \cong \mathfrak{B}(F)$. Пусть $\Phi_{f(n)}$ — аксиома для теории F . По построению имеем

$$\begin{aligned} n \in I &\rightarrow \Phi_{f(n)} \text{ полна,} \\ n \notin I &\rightarrow \Phi_{f(n)} \text{ не полна.} \end{aligned}$$

Теория $\Phi_{f(n)}$, как это отмечено в [8], будет всегда суперстабильной, если она полна. Тем более она будет стабильной. Отсюда получаем

$$\begin{aligned} n \in I &\rightarrow \Phi_{f(n)} \text{ полна и стабильна,} \\ n \notin I &\rightarrow \Phi_{f(n)} \text{ не полна.} \end{aligned}$$

Это даёт необходимую нижнюю оценку. \square

В заключение обсудим один открытый вопрос, естественно возникающий в связи с полученными результатами по конечно аксиоматизируемым теориям. Для этого нам понадобится дать несколько определений.

Интерпретацию I теории T в теории F назовём правильной, если выполнены условия

1. $\vdash_T \varphi \leftrightarrow \vdash_F \varphi^{(I)}$.
2. Множество $\{\varphi^{(I)} \mid \varphi \text{ — предложение теории } T\}$ порождает алгебру Линденбаума теории F .
3. $\varphi^{(I)}$ находится равномерно по φ .

Пусть C — свойство, которым могут обладать полные теории, I — интерпретация теории T в теории F , удовлетворяющая условиям 1, 2. Скажем, что I сохраняет свойство C , если для любого пополнения T^* теории T соответствующее пополнение F^* теории F , определяемое интерпретацией I , обладает свойством C тогда и только тогда, когда T^* обладает этим свойством.

Обозначим через K^* класс моделей языка

$$\mathcal{L}^* = \{P_0^1, P_1^1, \dots, P_n^1, \dots; n \in N\}$$

в которых нет конечных формульных множеств. Следующий результат представляет собой, в сущности, переформулировку теоремы 1 на языке интерпретаций.

ТЕОРЕМА 9. Каждая рекурсивно аксиоматизируемая теория T языка \mathcal{L}^* с условием $\text{Mod } T \subseteq K^*$ правильно интерпретируется в некоторой конечно аксиоматизируемой модельно полной теории F заданного конечного богатого языка \mathcal{L} с сохранением следующих свойств:

1. ω -стабильность.
2. Существование простой модели, её сильная конструктивизируемость и автоустойчивость.
3. Существование счётной насыщенной модели и её сильная конструктивизируемость.

Естественно поставить вопрос о справедливости подобного утверждения без ограничения на теорию и с большим списком свойств. Так приходим к весьма правдоподобной, как мне кажется, гипотезе о выразительных возможностях конечно аксиоматизируемых теорий:

ГИПОТЕЗА. Каждая рекурсивно аксиоматизируемая теория, не имеющая конечных моделей, правильно интерпретируется в некоторой конечно аксиоматизируемой теории с сохранением всех свойств из следующего списка:

1. Стабильность, суперстабильность, ω -стабильность.
2. Существование, сильная конструктивизируемость и автоустойчивость простой модели.
3. Существование, сильная конструктивизируемость и автоустойчивость минимальной модели.
4. Существование сильно конструктивной однородной модели.
5. Существование и сильная конструктивизируемость счётной насыщенной модели.
6. Модельная полнота.

ЛИТЕРАТУРА

1. С. С. Гончаров, *Некоторые свойства конструктивизаций булевых алгебр*, Сиб. матем. ж. 16, № 2 (1975), 264–278.
2. С. С. Гончаров и А. Т. Нуртазин, *Конструктивные модели полных разрешимых теорий*, Алгебра и логика 12, № 2 (1973), 125–142.
3. Ю. Л. Ершов, *Проблемы разрешимости и конструктивные модели*, Наука: Москва, 1980.
4. Г. Кейслер и Ч. Ч. Чэн, *Теория моделей*, Мир: Москва, 1977.
5. А. Т. Нуртазин, *Сильные и слабые конструктивизации и вычислимые семейства*, Алгебра и Логика 13, № 3 (1974), 311–323.
6. М. Г. Перетяткин, *Пример ω_1 -категоричной полной конечно аксиоматизируемой теории*, Алгебра и логика, 19, № 3 (1980), 314–347.
7. —, *Вычисления на машинах Тьюринга в конечно аксиоматизируемых теориях*, Алгебра и логика 21, № 4 (1982), 410–441.
8. —, *Конечно аксиоматизируемые тотально трансцендентные теории*, В кн.: Труды Института математики Сибирского отделения АН СССР, том 2, Математическая логика и теория алгоритмов, Наука: Новосибирск, 1982, с. 88–135.
9. Х. Роджерс, *Теория рекурсивных функций и эффективная вычислимость*, Мир: Москва, 1972.
10. L. Feiner, *Hierarchies of Boolean algebras*, J. Symbolic logic 35 (1970), 365–374.
11. W. Hanf, *Model theoretic methods in the study of elementary logic*, Symposium on the Theory of Models, North-Holland, Amsterdam, 1965, pp. 132–165.
12. —, *The Boolean algebra of logic*, Bull. Amer. Math. Soc. 31 (1975), 587–589.
13. L. Harrington, *Recursively presented prime models*, J. Symbolic Logic 39 (1974), 305–309.
14. A. H. Lachlan, *The transcendental rank of a theory*, Pacific J. Math. 27 (1971), 119–122.
15. R. L. Vaught, *Sentences true in all constructive models*, J. Symbolic Logic 25 (1960), 35–58.

Институт математики и механики, Академии наук казахской ССР, Алма-Ата 480100, СССР

On Schemes Axiomatizing Arithmetic

A. J. WILKIE

1. Introduction. It is commonly accepted that first-order Peano Arithmetic (PA) is the correct axiomatization of finitary mathematics. Of course this claim for PA cannot be mathematically proved but our aim here is to show that at least PA occupies a special position amongst a certain specific class of possible rivals. To describe this class and our results we introduce some notation.

Let L_1 (respectively L_2) denote the usual first- (respectively second-) order language of arithmetic. If $\Phi(X)$ is a formula of L_2 containing no set-quantifiers we may form the L_1 -scheme associated with Φ , denoted $\text{Scheme}(\Phi)$, which is the set of all L_1 -sentences of the form $\forall x_1, \dots, x_n \Phi(\psi)$, where $\psi = \psi(x_1, \dots, x_{n+1})$ is a formula of L_1 having no variables in common with Φ and $\Phi(\psi)$ is the result of replacing each subformula of Φ of the form " $x \in X$ " with $\psi(x_1, \dots, x_n, x)$.

Thus, for example, PA consists of a finite set of basic axioms, T_0 say, together with the theory $\text{Scheme}(I)$ where $I(X)$ is

$$(0 \in X \wedge \forall x(x \in X \rightarrow x + 1 \in X)) \rightarrow \forall x x \in X.$$

Now note that the sentence $\forall XI(X)$ is categorical over T_0 for the standard model $\mathbf{N} = \langle \omega; 0, 1, +, \cdot, \leq \rangle$. That is, for any L_1 -structure \mathcal{M} , $\mathcal{M} \models T_0 + \forall XI(X)$ if and only if $\mathcal{M} \cong \mathbf{N}$. (The set quantifier here is understood to range over all subsets of the domain.) It now seems natural to ask the question: Does PA occupy any special position amongst L_1 -theories of the form $\text{Scheme}(\Phi)$, where $\forall X\Phi(X)$ is categorical for \mathbf{N} over some finite L_1 -theory T ? In this generality I do not know an answer. However, if $\Phi(X)$ is of the following *restricted* form:

$$Q_1 x_1 \in X \cdots Q_k x_k \in X \phi(x_1, \dots, x_k)$$

(where ϕ is an L_1 -formula, the Q_i 's are \exists or \forall , and $\exists x \in X \cdots$, $\forall x \in X \cdots$ are abbreviations for $\exists x(x \in X \wedge \cdots)$, $\forall x(x \in X \rightarrow \cdots)$ respectively) then we can say something:

THEOREM 1.1. *Suppose $\Phi(X)$ is restricted and $\forall X\Phi(X)$ is categorical for \mathbf{N} (over some finite L_1 -theory T). Then there is a finite set, T_1 , of L_1 -sentences with $\mathbf{N} \models T_1$ such that $T_1 + \text{Scheme}(\Phi) \vdash \text{PA}$.*

REMARKS. (1) Let $W0(X)$ be the restricted formula:

$$\forall x_1 \in X \exists x_2 \in X \forall x_3 \in X x_2 \leq x_3.$$

Then, provided T_0 contains the usual successor axioms, $\forall XI(X)$ is equivalent to $\forall XW0(X)$ (over T_0) and $\text{Scheme}(I)$ is equivalent to $\text{Scheme}(W0)$ (over T_0). Thus PA may be regarded as a scheme associated with a categorical restricted formula and Theorem 1.1 is saying that, modulo finite true (in \mathbb{N}) L_1 -theories, it is the weakest such scheme.

(2) However, such schemes may be strictly stronger than PA. For example if we replace " \leq " in the formula $W0(X)$ by some formula naturally defining an ordering of ω of order type ε_0 (say), then it can be shown that for no finite T' (with $\mathbb{N} \models T'$) do we have $T' + \text{PA}$ proving all instances of the associated scheme.

(3) The "base-theory" T_1 required in Theorem 1.1 can be easily and explicitly determined from the formula Φ . Unfortunately, however, the theorem is perhaps devalued somewhat by the fact that it may be the case that $\text{PA} \not\vdash T_1$ (although it is known that for no finite consistent L_1 -theory, T' , do we have $T' \vdash \text{PA}$ —so the theorem does say something!). Indeed, we shall also prove the following result.

THEOREM 1.2. *There is a restricted formula $\Phi(X)$ such that $\forall X\Phi(X)$ is categorical for \mathbb{N} (over some finite PA-provable L_1 -theory T), but such that for no $n \in \omega$ do we have $I\Sigma_n + \text{Scheme}(\Phi) \vdash \text{PA}$, where $I\Sigma_n$ denotes PA with the induction scheme for Σ_n formulas only.*

Concerning our original question J. B. Paris has shown that Theorem 1.1 does not hold for categorical Π_1^1 sentences in general. Indeed, consider the formula, $\Psi(X)$, expressing the following:

" X codes a pair of sets $\langle Y, Z \rangle$ such that either $Y = \emptyset$ or Y has a least element or Z does not code a set of sentences of L_1 (with parameters) containing all true atomic and negated atomic sentences, no false ones, and satisfying the usual Tarski truth conditions."

Then it is easy to see that $\forall X\Psi(X)$ is categorical for \mathbb{N} (over some finite L_1 -theory). However, by Tarski's result on undefinability of truth we have that for any sufficiently strong, finite, true (in \mathbb{N}) L_1 -theory T , $T + \text{Scheme}(\Psi)$ is logically equivalent to $T!$

2. Approximating structures. Our proof of Theorem 1.1 uses a technique of approximating an infinite structure by a chain of finite ones, which is similar to Kripke's notion of fulfillment although our definitions here are taken from Shelah [5].

Let L be any finite relational language and $I = \langle I, < \rangle$ a totally ordered set. An I -structure is a sequence $\vec{A} = \langle A_i : i \in I \rangle$ of L -structures such that if $i, j \in I$ and $i < j$ then $A_i \subset A_j$. We write A_i for the domain of A_i and A for the domain of $\vec{A} = \bigcup_{\text{def.}} \vec{A}$. Let $L(A)$ be the language L together with a constant symbol \mathbf{a} for

each $a \in A$. If ϕ is a sentence of $L(A)$ in prenex normal form (p.n.f.), we define the notion $\vec{A} \Vdash \phi$ by induction on ϕ as follows:

- $$\begin{aligned} \vec{A} \Vdash \phi & \quad \text{iff } \mathcal{A} \models \phi \text{ for } \phi \text{ quantifier free,} \\ \vec{A} \Vdash \forall x \phi(x) & \quad \text{iff for all } a \in A, \vec{A} \Vdash \phi(\mathbf{a}), \\ \vec{A} \Vdash \exists x \phi(x) & \quad \text{iff whenever } i \in I \text{ is such that all the constant symbols, } \mathbf{a}, \text{ occurring in } \phi \text{ satisfy } a \in A_i, \text{ then for each } j > i \text{ there is some } b \in A_j \text{ such that } \vec{A} \Vdash \phi(\mathbf{b}). \end{aligned}$$

The following lemma is easily established by induction on the p.n.f. formula ϕ :

LEMMA 2.1. (i) Suppose I has no greatest element and $\vec{A} \Vdash \phi$. Then $\mathcal{A} \models \phi$.

(ii) Suppose $I' \subseteq I$. Let $\vec{A}' = \langle A_i : i \in I' \rangle$ and A' be the domain of $\bigcup \vec{A}'$. Then if $\phi \in L(A')$ and $\vec{A}' \Vdash \phi$, then $\vec{A} \Vdash \phi$. \square

It is convenient here to consider the class of formulas of $L(A)$ which are in p.n.f. and in which all negations are applied to atomic formulas. We denote this class by U . Note that every formula of $L(A)$ is logically equivalent to one in U . If $\neg\phi \in U$ we denote by ϕ^* the natural formula in U which is logically equivalent to ϕ (so that $\phi^{**} = \phi$), and we say a subset S of U is *closed* if S is closed under taking subformulas and if $\phi \in S$ implies $\phi^* \in S$. Notice that any finite $S \subseteq U$ is contained in some finite closed $\bar{S} \subseteq U$.

If $\phi(\vec{x}) \in U$ and \vec{A} is an I -structure, we say that \vec{A} *determines* $\phi(\vec{x})$ if whenever $\vec{a} \subseteq A$, then either $\vec{A} \Vdash \phi(\vec{a})$ or $\vec{A} \Vdash \phi(\mathbf{a})^*$. If $S \subseteq U$ we say that \vec{A} *determines* S if \vec{A} determines each formula in S .

LEMMA 2.2. Suppose $I = \langle \omega, < \rangle$ and \vec{A} is an I -structure with each A_i finite. Let $S \subseteq U$ be closed. Then there is an infinite increasing subsequence, $\langle i_j : j \in \omega \rangle$, of ω such that the I -structure $\langle A_{i_j} : j \in \omega \rangle$ determines S .

PROOF (by induction on the size of S). If S contains only quantifier-free formulas the result is clear. Otherwise let $\phi \in S$ be of maximal length. Then $S' = S \setminus \{\phi, \phi^*\}$ is clearly closed so by the inductive hypothesis we may as well suppose that \vec{A} determines S' .

We define the sequence $0 = i_0 < i_1 < \dots$ by induction such that for all $k \in \omega$ and $\vec{b} \subseteq A_{i_{k-1}}$, either $\vec{A}^{(k)} \Vdash \phi(\vec{b})$ or $\vec{A}^{(k)} \Vdash \phi(\vec{b})^*$, where $\vec{A}^{(k)} = \langle A_{i_0}, \dots, A_{i_k}, A_{1+i_k}, A_{2+i_k}, \dots \rangle$. The required result then follows from 2.1(ii).

So suppose i_0, \dots, i_k have been constructed and without loss of generality suppose $\phi(\vec{x})$ is $\exists y \phi'(y, \vec{x})$ where $\phi' \in S'$. For each $\vec{b} \subseteq A_{i_k} \setminus A_{i_{k-1}}$ (where we define $A_{i_{-1}} = \emptyset$) pick, if possible, $\nu(\vec{b}) \in A$ (where, as above, $A = \text{domain}(\bigcup \vec{A}) = \text{domain}(\bigcup \vec{A}^{(k)})$) such that $\vec{A}^{(k)} \Vdash \phi'(\nu(\vec{b}), \vec{b})$. Let i_{k+1} be the least $i > i_k$ such that for all $\vec{b} \subseteq A_{i_k} \setminus A_{i_{k-1}}$, if $\nu(\vec{b})$ is defined then $\nu(\vec{b}) \in A_i$.

Now suppose $\vec{b} \subseteq A_{i_k}$. We must show that either $\vec{A}^{(k+1)} \Vdash \phi(\vec{b})$ or $\vec{A}^{(k+1)} \Vdash \phi(\vec{b})^*$. If $\vec{b} \subseteq A_{i_{k-1}}$ this follows by the inductive hypothesis (for the construction). If $\vec{b} \subseteq A_{i_k} \setminus A_{i_{k-1}}$ and $\nu(\vec{b})$ is defined, we show $\vec{A}^{(k+1)} \Vdash \phi(\vec{b})$. Since

$\phi(\vec{b})$ contains a constant symbol, c say, such that $c \notin A_{i_{k-1}}$ (or, rather, since we may suppose this to be the case by the inductive hypothesis) it is sufficient to show that for some $d \in A_{i_{k+1}}$, $\vec{A}^{(k+1)} \models \phi'(\vec{d}, \vec{b})$. But this is clear (with $d = \nu(\vec{b})$) with k in place of $k+1$, and so the required conclusion follows from 2.1(ii). Now if $\vec{b} \subseteq A_{i_k} \setminus A_{i_{k-1}}$ and $\nu(\vec{b})$ is not defined, then for every $a \in A$, not $\vec{A}^{(k)} \models \phi'(\vec{a}, \vec{b})$. But $\phi' \in S'$ and $\vec{A}^{(k)}$ determines S' (by 2.1(ii), since \vec{A} determines S') so for every $a \in A$, $\vec{A}^{(k)} \models \phi'(\vec{a}, \vec{b})^*$, and hence, by definition, $\vec{A}^{(k)} \models \forall y(\phi'(y, \vec{b})^*)$, i.e., $\vec{A}^{(k)} \models \phi(\vec{b})^*$. But then $\vec{A}^{(k+1)} \models \phi(\vec{b})^*$ by 2.1(ii) and we are done. \square

3. The proof of Theorem 1.1. Let $\Phi(X) = Q_1x_1 \in X \cdots Q_kx_k \in X\phi(x_1, \dots, x_k)$ be a restricted formula, where $\phi(x_1, \dots, x_k) \in L_1$, and suppose that the L_2 -sentence $\forall X\Phi(X)$ is categorical for \mathbf{N} over some finite L_1 -theory T . ($\mathbf{N} \models T$ is implied by this.) Let L be the language containing just one k -ary relation symbol R , and let σ be the L -sentence $Q_1x_1 \cdots Q_kx_k R(x_1, \dots, x_k)$.

LEMMA 3.1. *For all $n \in \omega$ there is an $\langle n+1, < \rangle$ -structure $\vec{A} = \langle A_i : i \leq n \rangle$ (for L) such that for each $i \leq n$:*

- (i) A_i is finite.
- (ii) $A_i \subseteq \omega$ and $A_i \models R[n_1, \dots, n_k]$ iff $\mathbf{N} \models \phi[n_1, \dots, n_k]$ for all $n_1, \dots, n_k \in A_i$.
- (iii) $\vec{A} \models \sigma^*$.

PROOF. Let $M = \langle M; \dots \rangle$ be a countable nonstandard elementary extension of \mathbf{N} . Since $M \not\equiv \mathbf{N}$ there is some $S \subseteq M$ such that $M \models \neg\Phi[S]$, and we must have $S \neq \emptyset$ since $\Phi(\emptyset)$ can be written as a sentence of L_1 which is true in \mathbf{N} . (A similar argument, in fact, shows that S must be infinite.) Since S is countable we may choose (really) finite sets $B_0 \subseteq B_1 \subseteq \dots$ such that $S = \bigcup \{B_i : i \in \omega\}$. Let $V = \{\langle a_1, \dots, a_k \rangle \in M^k : M \models \phi[a_1, \dots, a_k]\}$ and let \mathcal{L}_i be the L -structure $\langle B_i; V \cap B_i^k \rangle$. By Lemma 2.2 (and preceding remarks) we may suppose the $\langle \omega, < \rangle$ -structure $\vec{\mathcal{L}} = \langle \mathcal{L}_i; i \in \omega \rangle$ determines σ . If $\vec{\mathcal{L}} \models \sigma$, then $\bigcup \vec{\mathcal{L}} \models \sigma$ by 2.1(i), which, since $\bigcup \vec{\mathcal{L}} = \langle S; V \cap S^k \rangle$, clearly implies $M \models \Phi[S]$ —a contradiction. Thus $\vec{\mathcal{L}} \not\models \sigma^*$ and hence, by 2.1(ii), for any $n \in \omega$, $\langle \mathcal{L}_i : i \leq n \rangle \models \sigma^*$. Now consider the following formula, $G(x)$ say, of L_1 :

$$\begin{aligned} & \text{"}\exists y, y \text{ is a (code for a) } \langle x+1, < \rangle\text{-structure such that } y \models \sigma^* \\ & \text{and } \forall a_1, \dots, a_k \in \bigcup y, y \models R[a_1, \dots, a_k] \leftrightarrow \phi(a_1, \dots, a_k)\text{"} \end{aligned}$$

We have shown that for all $n \in \omega$, $M \models G[n]$. Since $\mathbf{N} \preceq M$ the lemma follows. \square

To prove Theorem 1.1 we first define the base theory T_1 to be $I\Sigma_1 + \forall xG(x)$, where $G(x)$ is the formula from the above proof ($I\Sigma_1$ is known to be finitely axiomatizable). We have just shown that $\mathbf{N} \models \forall xG(x)$, and hence $\mathbf{N} \models T_1$, so it remains to show that $T_1 + \text{Scheme}(\Phi) \vdash \text{PA}$.

So suppose $\langle M; \dots \rangle = M \models T_1 + \text{Scheme}(\Phi)$ and

$$M \models \phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x+1)) \wedge \exists x \neg \phi(x),$$

where $\phi(x)$ is some formula of L_1 possibly involving parameters from M . Let $J = \{b \in M \mid M \models \forall x \leq b \phi(x)\}$. The J is a definable initial segment of M with no greatest element and we may choose $a \in M$ such that $a > b$ for all $b \in J$. Since $M \models G[a]$, there is a coded sequence $\vec{A} = \langle A_i : \forall i \leq a \rangle \in M$ of L -structures such that

$$\begin{aligned} M \models & \text{"}\forall i \leq j \leq a, A_i \subseteq A_j \text{ and } \vec{A} \Vdash \sigma^* \text{ and} \\ & \forall a_i, \dots, a_k \in \text{domain}(A_i), A_i \models R[a_1, \dots, a_k] \leftrightarrow \phi(a_1, \dots, a_k)\text{"} \end{aligned} \quad (*)$$

Since M satisfies a sufficiently strong fragment of PA (namely $I\Sigma_1$) the sequence \vec{A} may be regarded as an I_a -structure in the real world such that $\vec{A} \Vdash \sigma^*$ (where $I_a = \langle \{b \in M : M \models b \leq a\}, <^M \rangle$). It follows from 2.1(ii) that $\langle A_i : i \in J \rangle \Vdash \sigma^*$ and hence, by 2.1(i), that $\bigcup \langle A_i : i \in J \rangle \models \sigma^*$. Now the domain, A say, of this union is defined in M by the formula $\exists y (\forall z \leq y \phi(z) \wedge x \in \text{domain}(A_y)) = \chi(x)$ say, and, by (*), the interpretation of R is exactly $A^k \cap \{(a_1, \dots, a_k) \in M^k : M \models \phi[a_1, \dots, a_k]\}$. It follows that M satisfies the sentence $Q_1 x_1 \cdots Q_k x_k \phi(x_1, \dots, x_k)$ with the $Q_i x_i$'s relativized to the formula χ , which violates an instance of Scheme(Φ) and establishes Theorem 1.1.

4. The proof of Theorem 1.2. Although we have not mentioned indiscernibles so far, the reader may have noticed that they are implicitly at work in the above. Indeed, Lemma 2.2 is just a version of Ramsey's theorem. It is therefore not surprising that to prove Theorem 1.2 we shall construct $\Phi(X)$ so that any finite T_1 which brings Scheme(Φ) up to the strength of PA must itself have roughly the strength of the Paris-Harrington independent statement. We refer the reader to [2, 3, or 4] for the facts on Paris-Harrington theory that we shall use, without proof, below.

We begin, then, by letting $\theta(x, y)$ be a Σ_0 (i.e., bounded) formula of L_1 such that

$$I\Sigma_1 \vdash \forall x, y (\theta(x, y) \rightarrow (y > 2x \wedge [x, y] \rightarrow (*)_x^x)), \quad (1)$$

and

$$I\Sigma_1 \vdash \forall x, z ((z > 2x \wedge [x, z] \rightarrow (*)_x^x) \rightarrow \exists y \theta(x, y)), \quad (2)$$

where $[x, y] \rightarrow (*)_x^z$ is the usual Paris-Harrington combinatorial relation. (Such a formula θ is easily constructed by "padding" this relation.) We now define $\Phi(X)$ to be the restricted formula:

$$\forall v \in X \exists z \in X \forall x \in X (\forall y (\theta(x, y) \rightarrow z \leq y)).$$

Then $N \models \forall X \Phi(X)$, since if $S \subseteq \omega$ then either $S = \emptyset$ in which case $N \models \Phi[S]$ holds vacuously, or else we may choose z to be the least element of S and observe that $N \models I\Sigma_1 \vdash \forall x, y (\theta(x, y) \rightarrow y > x)$ (by (1)).

To see that $\forall X \Phi(X)$ is categorical for N over, say, $I\Sigma_1$, suppose $\langle M; \dots \rangle = M \models I\Sigma_1 + \forall X \Phi(X)$ and that M is nonstandard. Let $S = M \setminus \omega$. Then $S \neq \emptyset$ so for some $a \in S$, $M \models \forall x \in S \forall y (\theta(x, y) \rightarrow a \leq y) \cdots (*)$. However, for each $n \in \omega$ we have (by (1), (2)) that $N \models \exists y \theta(n, y)$ and so (since θ is Σ_0),

$\mathcal{M} \models \exists y(\theta(n, y) \wedge y < a)$. Since $\mathcal{M} \models I\Sigma_1$ (in fact $\mathcal{M} \models I\Sigma_0$ is sufficient here) we have for some nonstandard $b \in M$, $\mathcal{M} \models \exists y(\theta(b, y) \wedge y < a)$ which contradicts (*). Thus $\mathcal{M} \cong \mathbb{N}$ as required.

We shall now complete the proof of Theorem 1.2 by showing that for no $n \in \omega$ do we have $I\Sigma_n + \text{Scheme}(\Phi) \vdash \text{PA}$. So suppose this false for some $n \in \omega$, $n \geq 1$. Let $k = 2n + 4$. Since $\text{PA} \vdash \forall x \exists y [x, y] \rightarrow (*)_x^k$, there is some $m \in \omega$ such that

$$I\Sigma_n + \text{Scheme}_m(\Phi) \vdash \forall x \exists y [x, y] \rightarrow (*)_x^k, \quad (3)$$

where $\text{Scheme}_m(\Phi)$ denotes the set of all sentences of the form $\forall \vec{x} \Phi(\psi)$ where ψ is a Σ_m formula.

Let $\mathcal{M} = \langle M; \dots \rangle$ be a countable nonstandard elementary extension of \mathbb{N} and pick $a \in M \setminus \omega$. Let $F(x) = \mu y : [x, y] \rightarrow (*)_x^{k-2}$ and define \hat{F} by $\hat{F}(0, x) = x$, $\hat{F}(y + 1, x) = F(\hat{F}(y, x))$. Define, for $n \in \omega$, $l_0(x) = x$, $l_{n+1}(x) = 2^{l_n(x)}$. Then these are all \mathcal{M} -definable functions (which are total in \mathcal{M}) and by a result of [6] (Theorem 11, p. 338) there is an $r \in \omega$ such that if $\beta \in M$ is chosen to satisfy $l_r(\beta) < a \leq l_r(\beta + 1)$, then there is an initial segment substructure $\mathcal{A} = \langle A; \dots \rangle$ of \mathcal{M} such that:

$$a \in A \text{ and } \hat{F}(a, a) \notin A. \quad (4)$$

$$\begin{aligned} A \text{ is closed under } F \text{ and for all } b, c, d, e \in A, \mathcal{A} \models [b, c] \rightarrow (*)_e^d \text{ iff} \\ \mathcal{M} \models [b, c] \rightarrow (*)_e^d. \end{aligned} \quad (5)$$

$$\begin{aligned} \text{If } \psi(x) \text{ is any } \Sigma_{n+m+3} \text{ formula of } L_1(\text{with parameters from } A), \\ \text{then } \mathcal{A} \models (\psi(0) \wedge \forall x(\psi(x) \rightarrow \psi(x+1))) \rightarrow \psi(\beta). \end{aligned} \quad (6)$$

We now consider two cases and derive a contradiction in each case which will establish Theorem 1.2.

Case 1. $\mathcal{A} \models I\Sigma_n$. It is known that the least $y \in M$ satisfying $\mathcal{M} \models [a, y] \rightarrow (*)_a^k$ is (much) larger than $\hat{F}(a, a)$. Hence, by (4), (5), $\mathcal{A} \models \neg \exists y [a, y] \rightarrow (*)_a^k$. Thus, we contradict (3) if we show $\mathcal{A} \models \text{Scheme}_m(\Phi)$.

So suppose $\phi(x)$ is a Σ_m formula (with parameters from A) and $\mathcal{A} \models \exists x \phi(x)$. We must show that

$$\mathcal{A} \models \exists z(\phi(z) \wedge \forall x(\phi(x) \rightarrow \forall y(\theta(x, y) \rightarrow z \leq y))). \quad (7)$$

Now if there is a least $b \in A$ such that $\mathcal{A} \models \phi[b]$, then (7) is immediate by (1). Otherwise we have $\mathcal{A} \models \psi(0) \wedge \forall x(\psi(x) \rightarrow \psi(x+1))$, where $\psi(x)$ is $\forall y \leq x \neg \phi(y)$, and hence, by (6), $\mathcal{A} \models \psi[\beta]$. In this case we shall show that $\mathcal{A} \models \forall x(\phi(x) \rightarrow \forall y \neg \theta(x, y))$, which certainly establishes (7). So suppose, for contradiction, that $d, e \in A$ and $\mathcal{A} \models \phi(d) \wedge \theta(d, e)$. Since $\mathcal{A} \models \forall y \leq \beta \neg \phi(y)$ we have $d > \beta$ and, by (1), $\mathcal{A} \models e > 2d \wedge [d, e] \rightarrow (*)_d^d$. This implies (using (5)) that $\mathcal{M} \models 2d < e < \hat{F}(a, a) \wedge [\beta, e] \rightarrow (*)_\beta^\beta$. However, the least $y \in M$ such that $\mathcal{M} \models [\beta, y] \rightarrow (*)_\beta^\beta$ is (since β is nonstandard) known to be much larger than $\max\{\hat{F}(t, t) : t \leq l_r(\beta + 1)\} (\geq \hat{F}(a, a))$.

Case 2. $\mathcal{A} \not\models I\Sigma_n$. This implies there is a Σ_{n+1} formula (with parameters from A) which defines in \mathcal{A} a proper initial segment, J say, with no greatest element. Choose $c \in A \setminus J$. Then, by (6), $\beta < c$. By (5) there is a minimal $d \in M$ such

that $d \in A$ and $M \models [c, d] \rightarrow (*)_c^{k-2}$. Now (working in M) there is, by Paris-Harrington theory (using $k - 2 = 2n + 2$), a coded sequence $\vec{e} = \langle e_1, \dots, e_c \rangle \in [c, d]^c$ such that the substructure \mathcal{L} say, of M with domain $B = \{a \in M : a \leq e_i \text{ for some } i \in J\}$ is an initial segment of M satisfying $I\Sigma_n$. Now (5) certainly implies $\vec{e} \in A$, so B is definable in \mathcal{A} by a Σ_{n+1} formula (using the parameter \vec{e} and those needed to define J). Further, the relation $[x, y] \rightarrow (*)_t^z$ is absolute between \mathcal{L}, \mathcal{A} , and M and so, since $d \notin B$, we have that $\mathcal{L} \models \neg \exists y [c, y] \rightarrow (*)_c^{k-2}$. Thus (3) will be contradicted if we show $\mathcal{L} \models \text{Scheme}_m(\Phi)$. So suppose $\phi(x)$ is a Σ_m formula (with parameters from B) such that $\mathcal{L} \models \exists x \phi(x)$. We must show

$$\mathcal{L} \models \exists z (\phi(z) \wedge \forall x (\phi(x) \rightarrow \forall y (\theta(x, y) \rightarrow z \leq y))). \quad (8)$$

As above, this is clear if there is a minimal $b \in B$ such that $\mathcal{L} \models \phi[b]$, so we may suppose the Σ_{m+1} formula $\psi(x) \stackrel{\text{def.}}{=} \forall y \leq x \neg \phi(y)$ defines in \mathcal{L} an initial segment, K say, with no greatest element. Now recall that B is definable in \mathcal{A} by a Σ_{n+1} formula, so K is definable in \mathcal{A} by, certainly, a Σ_{n+m+3} formula and hence, by (6), $\beta \in K$. The proof of (8) is now similar to that of (7) in Case 1.

ACKNOWLEDGMENT. I wish to express my thanks to Dan Isaacson for many discussions on the content of this paper. In particular it was his article [1] (which discusses the first sentence of this paper) which motivated me to consider the questions I have been investigating here.

REFERENCES

1. Daniel Isaacson, *Arithmetical truth and hidden higher-order concepts*, Proc. 1985 Summer European Meeting of the Assoc. for Symbolic Logic at Pairs (Orsay), North-Holland (to appear).
2. Jussi Ketonen and Robert Solovay, *Rapidly growing Ramsey functions*, Ann. of Math. (2) **113** (1981), 267–314.
3. J. B. Paris, *A hierarchy of cuts in models of arithmetic*, Model Theory of Algebra and Arithmetic, Lecture Notes in Math., vol. 834, Springer-Verlag, 1980, pp. 312–337.
4. J. B. Paris and L. Harrington, *A mathematical incompleteness in Peano arithmetic*, Handbook of Mathematical Logic, J. Barwise, editor, North-Holland, 1977.
5. Saharon Shelah, *On logical sentences in PA*, Logic Colloquium '82 (Florence 1982), North-Holland, 1984, pp. 145–160.
6. A. J. Wilkie, *On sentences interpretable in systems of arithmetic*, Logic Colloquium '84, J. B. Paris, A. J. Wilkie, and G. M. Wilmers, editors, North-Holland, 1986.

MATHEMATICAL INSTITUTE, OXFORD OX1 3LB, ENGLAND

The What, Where, and Why of Almost Split Sequences

MAURICE AUSLANDER

Throughout this paper we assume that R is a noetherian, local Gorenstein ring. We also assume that all R -algebras Λ are finitely generated R -modules. An R -module Λ is said to be complete if R is complete and is said to be Cohen-Macaulay if Λ is a (maximal) Cohen-Macaulay R -module. Recall that if Λ is a complete R -algebra, then $\text{mod } \Lambda$, the category of finitely generated Λ -modules, is a Krull-Schmidt category; i.e., an M in $\text{mod } \Lambda$ is indecomposable if and only if $\text{End}_\Lambda(M)$, the endomorphism ring of M , is a local ring and thus the decomposition of an X in $\text{mod } \Lambda$ into a sum (direct) of indecomposables is unique up to isomorphism. When Λ is a Cohen-Macaulay R -algebra, we denote by $\text{CM } \Lambda$ the full subcategory of $\text{mod } \Lambda$ consisting of the X in $\text{mod } \Lambda$ which are (maximal) Cohen-Macaulay modules when viewed as R -modules. Further, if Λ is a Cohen-Macaulay R -algebra we have the duality $D: \text{CM } \Lambda \rightarrow \text{CM } \Lambda^{\text{op}}$ given by $D(X) = \text{Hom}_R(X, R)$ for all X in $\text{CM } \Lambda$. Moreover, this duality is exact in the sense that if $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ is exact in $\text{CM } \Lambda$, i.e., it is an exact sequence of Λ -modules with A, B, C in $\text{CM } \Lambda$, then $0 \rightarrow D(C) \rightarrow D(B) \rightarrow D(A) \rightarrow 0$ is exact in $\text{CM } \Lambda^{\text{op}}$. We say that A in $\text{CM } \Lambda$ is $\text{CM } \Lambda$ -injective if every exact sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\text{CM } \Lambda$ splits. From the exact duality $D: \text{CM } \Lambda \rightarrow \text{CM } \Lambda^{\text{op}}$ it follows that A is $\text{CM } \Lambda$ -injective if and only if $A \cong D(P)$ for some projective Λ^{op} -module.

We will be mainly interested in questions concerning almost split sequences in $\text{CM } \Lambda$ when Λ is a complete Cohen-Macaulay R -algebra. Recall that an almost split sequence in $\text{CM } \Lambda$ is a nonsplit exact sequence $0 \rightarrow A \xrightarrow{f} B \xrightarrow{g} C \rightarrow 0$ in $\text{CM } \Lambda$ having the following additional properties: (a) A and C are indecomposable; (b) if $h: X \rightarrow C$ is a nonisomorphism in $\text{CM } \Lambda$ with X indecomposable, then h can be lifted to B ; i.e., there is a $t: X \rightarrow B$ such that $gt = h$; (c) each nonisomorphism $j: A \rightarrow Y$ in $\text{CM } \Lambda$ with Y indecomposable can be extended to B ; i.e., there is an $s: B \rightarrow Y$ such that $sf = j$. It should be observed that (a) and (b) is equivalent to (a) and (c) and that two almost split sequences

Written with partial support from National Science Foundation Grant MCS 83-03348.

$0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ and $0 \rightarrow A' \rightarrow B' \rightarrow C' \rightarrow 0$ in $\text{CM } \Lambda$ are isomorphic if and only if $A \approx A'$ if and only if $C \cong C'$.

Before stating our main existence theorem for almost split sequences in $\text{CM } \Lambda$, it is convenient to make the following definition. We say that a Λ -module X has a certain property outside the maximal ideal if the $\Lambda_{\mathfrak{p}}$ -module $X_{\mathfrak{p}}$ has the property for all nonmaximal ideals \mathfrak{p} of R .

THEOREM 1 (AUSLANDER-REITEN [8]). *Suppose Λ is a complete Cohen-Macaulay R -algebra.*

(a) *An indecomposable nonprojective C in $\text{CM } \Lambda$ has the property that it is projective outside the maximal ideal if and only if there is an almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\text{CM } \Lambda$.*

(b) *An indecomposable A in $\text{CM } \Lambda$ which is not $\text{CM } \Lambda$ -injective is $\text{CM } \Lambda$ -injective outside the maximal ideal if and only if there is an almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\text{CM } \Lambda$.*

Let $P_m(\text{CM } \Lambda)$ and $I_m(\text{CM } \Lambda)$ denote the full subcategories of $\text{CM } \Lambda$ consisting of the X which are projective ($\text{CM } \Lambda$ -injective) outside the maximal ideal and have no projective ($\text{CM } \Lambda$ -injective) indecomposable summands. Combining Theorem 1 with the fact that either end of an almost split sequence determines the sequence (up to isomorphism), we see that we obtain the following bijection between the isomorphism classes of indecomposable objects in $P_m(\text{CM } \Lambda)$ and $I_m(\text{CM } \Lambda)$. We send $[C]$, the isomorphism class of an indecomposable C in $P_m(\text{CM } \Lambda)$, to the uniquely determined $[A]$ with A indecomposable in $I_m(\text{CM } \Lambda)$ such that there is an almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$. We now give a description of this bijection without reference to almost split sequences (Auslander [4]).

Let $P_1 \xrightarrow{f} P_0 \rightarrow C \rightarrow 0$ be a minimal projective presentation of C in $P_m(\text{CM } \Lambda)$. As usual, define $\text{Tr } C$, the transpose of C , by the exact sequence $P_0^* \xrightarrow{f^*} P_1^* \rightarrow \text{Tr } C \rightarrow 0$ where $X^* = \text{Hom}_{\Lambda}(X, \Lambda)$. Then define $\text{Tr}_L C = \Omega^d \text{Tr } C$ where Ω^d is the d th syzygy in a minimal projective resolution of $\text{Tr } C$ and dimension R is d . Then $\text{Tr}_L C$ is in $P_m(\text{CM } \Lambda^{\text{op}})$ and $D \text{Tr}_L C$ is in $I_m(\text{CM } \Lambda)$. Moreover, if $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ is an almost split sequence in $\text{CM } \Lambda$, then $A \cong D \text{Tr}_L C$ and $C \cong \text{Tr}_L D A$.

As a consequence of Theorem 1, we have the following characterization of when a complete Cohen-Macaulay R -algebra Λ is an isolated singularity; i.e., $\text{gl dim } \Lambda_{\mathfrak{p}} = \dim R_{\mathfrak{p}}$ for all nonmaximal ideals \mathfrak{p} of R .

THEOREM 2 (AUSLANDER [5, 4]). *The following are equivalent for a complete, Cohen-Macaulay R -algebra Λ .*

(a) *Λ is an isolated singularity.*

(b) $P_m(\Lambda) = \text{CM } \Lambda = I_m(\Lambda)$.

(c) *For each indecomposable, nonprojective C in $\text{CM } \Lambda$ there is an almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\text{CM } \Lambda$.*

Recall that a complete, Cohen-Macaulay R -algebra is said to be of finite Cohen-Macaulay type if $\text{CM } \Lambda$ has only a finite number of nonisomorphic indecomposable modules. Now it is not difficult to see that if Λ is a complete Cohen-Macaulay R -algebra of finite Cohen-Macaulay type, then every nonprojective C in $\text{CM } \Lambda$ has an almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\text{CM } \Lambda$. Therefore we have the following result as an immediate consequence of Theorem 2.

PROPOSITION 3 (AUSLANDER [3, 5]). *A complete, Cohen-Macaulay R -algebra is an isolated singularity if it is of finite Cohen-Macaulay type.*

An interesting question is to classify the complete Cohen-Macaulay rings of the form $\Lambda = k[[X_1, \dots, X_d]]/\mathfrak{a}$ which are of finite Cohen-Macaulay type, where \mathfrak{a} is an ideal in the formal power series ring $k[[X_1, \dots, X_d]]$ over an algebraically closed field k . This has been accomplished by Knörrer [24] and Buchweitz, Greuel, and Schreyer [17], when \mathfrak{a} is a principal ideal and characteristic $k \neq 2$. When $\text{ch } k = 0$, one obtains precisely the Arnold simple singularities. When $k = \mathbb{C}$, the complex numbers, and dimension $\Lambda = 2$, then Herzog [23] has shown that the quotient singularities with respect to finite subgroups of $\text{GL}(2, \mathbb{C})$ are of finite Cohen-Macaulay type while Artin-Verdier (unpublished), Auslander [6], and Ésnault [19] have shown the converse. Auslander-Reiten [12] have found the only known nonhypersurface singularities of finite Cohen-Macaulay type in dimensions greater than 2, both of which have dimension 3. They are the fixed ring $\mathbb{C}[[X_1, X_2, X_3]]^{Z/2Z}$ where the generator σ of $Z/2Z$ operates by $\sigma(X_i) = -X_i$ for all i , which is the only quotient singularity of dimension greater than 2 of finite Cohen-Macaulay type, and the scroll

$$\mathbb{C}[[X_0, X_1, X_2, Y_0, Y_1]]/(X_0X_2 - X_1^2, X_0X_1 - X_1Y_0, X_1Y_1 - X_2Y_0).$$

In this connection, it should be noted that Artin [1] has classified an interesting class of noncommutative complete Cohen-Macaulay R -algebras Λ of finite Cohen-Macaulay type. Namely, when $R = k[[X_1, X_2]]$, $\text{Ch } k = 0$ and Λ is a maximal order in a division ring of finite dimension over the field of quotients of R .

Now when $\dim R = 0$, then a complete, Cohen-Macaulay R -algebra Λ is nothing more than an R -algebra Λ and $\text{mod } \Lambda = \text{CM } \Lambda = P_m(\text{CM } \Lambda) = I_m(\text{CM } \Lambda)$. For example, when R is a field, one is dealing with the theory of finite-dimensional modules over a finite-dimensional algebra. So under these circumstances Theorem 1 says that every indecomposable nonprojective Λ -module C has an almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\text{mod } \Lambda$. While this now appears as a special case of Theorem 1, it was in the zero-dimensional case that almost split sequences were first introduced by Auslander-Reiten [13] and where they have been most extensively used, to very good effect. For instance, the first examples of tiltings [7] and coverings [26] depended in an essential way on almost split sequences. Since then tiltings and coverings as developed by Brenner-Butler [16], Happel-Ringel [22], and Bongartz-Gabriel [15] respectively have become

major tools in studying the representation theory of finite-dimensional algebras. The rest of this paper is devoted to showing how some ideas and results first developed in the zero-dimensional case have appeared in higher dimensions.

From now on we assume that Λ is a complete Cohen-Macaulay R -algebra which is a commutative local ring. Since every commutative complete Cohen-Macaulay local ring S is a complete Cohen-Macaulay R -algebra for some R , we are really just talking about commutative complete Cohen-Macaulay rings S . To emphasize this point, we will denote Λ by S and often drop the R . In addition we assume that S is an isolated singularity which simply means that $S_{\mathfrak{p}}$ is regular for all nonmaximal prime ideals of S . Therefore $\text{CM } S = P_{\mathfrak{m}}(\text{CM } S) = I_{\mathfrak{m}}(\text{CM } S)$ so that there is an almost split sequence $0 \rightarrow A \xrightarrow{g} B \xrightarrow{f} C \rightarrow 0$ in $\text{CM } S$ for each indecomposable $C \not\cong S$ and for each $A \not\cong \text{Hom}_R(S, R)$.

Let C be an indecomposable module in $\text{CM } S$. Then a morphism $f: B \rightarrow C$ in $\text{CM } S$ is said to be a minimal right almost split morphism if it satisfies the following: (a) f is not a splittable surjection; (b) if $t: X \rightarrow C$ in $\text{CM } S$ is not a splittable surjection, then it can be lifted to B ; (c) if $h: B \rightarrow B$ is such that $fh = f$, then h is an automorphism. It is easily seen that if $f: B \rightarrow C$ and $f': B' \rightarrow C$ are two minimal right almost split morphisms in $\text{CM } S$, then there is an isomorphism $h: B \rightarrow B'$ such that $f = f'h$, so the minimal right almost split morphisms to C are uniquely determined (up to isomorphism) by C . If $C \not\cong S$ and $0 \rightarrow A \xrightarrow{g} B \xrightarrow{f} C \rightarrow 0$ is an almost split sequence in $\text{CM } S$, then $f: B \rightarrow C$ is easily seen to be a minimal right almost split morphism. If $C \cong S$, then it can be shown that there is also a minimal right almost split morphism $E \rightarrow S$ in $\text{CM } S$. So for each indecomposable C in $\text{CM } S$ there is a uniquely determined (up to isomorphism) minimal right almost split morphism $f: B \rightarrow C$. Also we have the dual notion of a minimal left almost split morphism $g: A \rightarrow B$ in $\text{CM } S$ with A indecomposable and the dual result that given any indecomposable A in $\text{CM } S$, there is a uniquely determined (up to isomorphism) minimal left almost split morphism $A \xrightarrow{g} B$ which is the left-hand side of the almost split sequence $0 \rightarrow A \xrightarrow{g} B \xrightarrow{f} C \rightarrow 0$ in $\text{CM } S$ when $A \not\cong \text{Hom}_R(S, R)$.

A notion intimately connected to that of minimal right (left) almost split morphisms is that of an irreducible morphism first introduced in the case of Artin algebras by Auslander-Reiten [14]. A morphism $f: X \rightarrow Y$ between indecomposable modules in $\text{CM } S$ is said to be irreducible in $\text{CM } S$ if (a) f is not an isomorphism and (b) if $f = hg$ in $\text{CM } S$, then g is a splittable injection or h is a splittable surjection. Let X and Y be indecomposable and suppose $g: X \rightarrow B$ and $h: B' \rightarrow Y$ are minimal left and right almost split morphisms respectively. Then for $f: X \rightarrow Y$ in $\text{CM } S$, the following are equivalent: (a) f is irreducible; (b) there exists splittable injection $s: X \rightarrow B'$ such that $hs = f$; (c) there exists a splittable surjection $t: B \rightarrow Y$ such that $tg = f$. Thus for each indecomposable module X in $\text{CM } S$, there exist only a finite number of nonisomorphic indecomposable Y in $\text{CM } S$ which have either an irreducible morphism $Y \rightarrow X$ or an irreducible morphism $X \rightarrow Y$.

As our first example of how irreducible morphisms have been used in arbitrary dimensions we state the following criterion for determining when S is of finite Cohen-Macaulay type which is essentially the same as a criterion used extensively in the zero-dimensional case [4]. This criterion was used to show that the scroll $k[[X_0, X_1, X_2, Y_0, Y_1]]/(X_0X_2 - X_1^2, X_0X_1 - X_1Y_0, X_1Y_1 - X_2Y_0)$, cited earlier, is of finite Cohen-Macaulay type.

PROPOSITION 4 (AUSLANDER-REITEN [12]). *Suppose \mathcal{C} is a subcategory of $\text{CM } S$ closed under isomorphisms consisting of a finite number of nonisomorphic indecomposable modules including S . Further, assume that an indecomposable X in $\text{CM } S$ is in \mathcal{C} if there exists an irreducible $C \rightarrow X$ or an irreducible $X \rightarrow C$ with C in \mathcal{C} . Then S is of finite Cohen-Macaulay type and \mathcal{C} consists of all the indecomposable modules in $\text{CM } S$.*

Assume now that S is a k -algebra where k is an algebraically closed field and $S/\mathfrak{m} = k$ where \mathfrak{m} is the maximal ideal of S . Then if $U \rightarrow Y$ and $X \rightarrow V$ are minimal right and left almost split morphisms in $\text{CM } S$ with Y and X indecomposable, then the multiplicity that Y appears as a summand of V is the same as the multiplicity X appears as a summand of U . This common number is called the number of independent irreducible maps from X to Y . We are now in position to define the AR-quiver of S . This is a directed graph whose vertices are the isomorphism classes $[Z]$ of indecomposable modules Z in $\text{CM } S$ and where the number of arrows from $[X]$ to $[Y]$ is the number of independent irreducible morphisms from X to Y . When S is a Gorenstein ring, the AR-quiver of S minus the vertex $[S]$ and the arrows to it and out of it, is called the stable AR-quiver of S . We then have the following striking result which is the consequence of work of Gonzalez-Springberg and Verdier [21], Artin-Verdier [2], Ésnault-Knörrer [20], Auslander-Reiten [9].

THEOREM 5. *Suppose S is a rational double point over an algebraically closed field of any characteristic. Then the desingularization graph of S is naturally isomorphic to the underlying graph of the stable AR-quiver of S .*

In connection with this theorem it is of interest to point out a different but related interpretation of the AR-quiver of two-dimensional quotient singularities in terms of finite group representations. This requires the notion of the McKay-quiver of a representation of a finite group.

Let k be an algebraically closed field and G a finite group whose order is not divisible by $\text{Ch } k$. Let V_1, \dots, V_s be a complete list of nonisomorphic simple kG -modules. Let V be an arbitrary (finite-dimensional) $k[G]$ -module. Then the McKay-quiver of V is defined as follows. The V_i are the vertices and one draws n arrows from V_i to V_j if V_i occurs in $V \otimes_k V_j$ with multiplicity n .

Suppose now that V is a faithful two-dimensional kG -module. Associated with this is a linear action of G as a group of k -automorphisms of $k[[X_1, X_2]]$ whose fixed ring we denote by S . Then S is a local complete integrally closed two-dimensional domain and therefore an isolated Cohen-Macaulay singularity.

In addition, assume that the kG -module V has the property that the height one primes of S are unramified in $k[[X_1, X_2]]$ (if $k = \mathbf{C}$, the complex numbers, this means that G has no pseudo-reflections). Then, as noted before, S is of finite Cohen-Macaulay type and we have

PROPOSITION 6 (AUSLANDER [6]). *Under the above hypothesis there is a natural isomorphism between the McKay-quiver of V and the AR-quiver of S .*

Since McKay [25] had already observed that for G a finite subgroup of $\mathrm{SL}(2, \mathbf{C})$, the McKay-quiver of V , the kG -module given by the inclusion $G \subset \mathrm{SL}(2, \mathbf{C})$, is isomorphic to the desingularization graph of the associated fixed ring, Proposition 6 gave the first connection between almost split sequences and singularity theory.

As our final remark concerning two-dimensional rings, we point out some special features of the structure of almost split sequences for arbitrary S of dimension 2 which played a significant role in the proofs of the above results.

Suppose $\dim S = 2$ and let $\omega = \mathrm{Hom}_R(S, R)$, the dualizing module of S . Then there is a unique (up to isomorphism) exact sequence $0 \rightarrow \omega \xrightarrow{g} E \xrightarrow{f} S \rightarrow S/\mathfrak{m} \rightarrow 0$ with E in $\mathrm{CM} S$ called the fundamental sequence of S . It is not difficult to see that $f: E \rightarrow S$ is the minimal right almost split morphism to S .

PROPOSITION 6 (AUSLANDER [6] AND AUSLANDER-REITEN [10]). *Suppose $\dim S = 2$ and $0 \rightarrow \omega \xrightarrow{g} E \xrightarrow{f} S \rightarrow S/\mathfrak{m} \rightarrow 0$ is the fundamental sequence and let $M \not\cong S$ be an indecomposable module in $\mathrm{CM} S$ and let $M^* = \mathrm{Hom}_S(M, S)$. Then the induced exact sequence $0 \rightarrow \mathrm{Hom}_S(M^*, \omega) \rightarrow \mathrm{Hom}_S(M^*, E) \rightarrow \mathrm{Hom}_S(M^*, S) = M \rightarrow 0$ has the following properties.*

(a) $0 \rightarrow \mathrm{Hom}_S(M^*, \omega) \rightarrow \mathrm{Hom}_S(M^*, E) \rightarrow M \rightarrow 0$ is exact and is either split or almost split.

(b) If $\mathrm{ch} S/\mathfrak{m}$ does not divide $\mathrm{rank} M$, then

$$0 \rightarrow \mathrm{Hom}_S(M^*, \omega) \rightarrow \mathrm{Hom}_S(M^*, E) \rightarrow M \rightarrow 0$$

is almost split.

(c) If S/\mathfrak{m} is algebraically closed, then

$$0 \rightarrow \mathrm{Hom}_S(M^*, \omega) \rightarrow \mathrm{Hom}_S(M^*, E) \rightarrow M \rightarrow 0$$

is almost split if and only if $\mathrm{ch} S/\mathfrak{m}$ does not divide $\mathrm{rank} M$.

We next turn our attention to computing $G_0(S)$, the Grothendieck group of S , which is the free abelian group with basis the isomorphism classes $[X]$ of X in $\mathrm{mod} S$ modulo the subgroup generated by $[A] - [B] + [C]$ whenever there is an exact sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\mathrm{mod} S$. Because S is a Cohen-Macaulay ring, it is easily seen that $G_0(S)$ is also the free abelian group with basis $[X]$ with X in $\mathrm{CM} S$ modulo the subgroup generated by $[A] - [B] + [C]$ whenever there is an exact sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$ in $\mathrm{CM} S$. While the split and almost split sequences in $\mathrm{CM} S$ give some of the relations for $G_0(S)$ in general, it is when S is of finite Cohen-Macaulay type that we get a really definitive

connection between Grothendieck groups and almost split sequences which is a generalization of Butler's result [18] in dimensions at most 1.

PROPOSITION 6 (AUSLANDER-REITEN [11]). *Suppose S is of finite Cohen-Macaulay type. Then $G_0(S)$ is the free abelian group with basis the isomorphism classes of X in $\text{CM } S$ modulo the subgroup generated by $[A] - [B] + [C]$ whenever there is a split or almost split sequence $0 \rightarrow A \rightarrow B \rightarrow C \rightarrow 0$.*

As an easy consequence of this result we have

COROLLARY 7. *Let S be of finite Cohen-Macaulay type. Then $G_0(S)$ is completely determined by the AR-quiver of S minus the vertex $[S]$ and all arrows going into and coming out of it.*

We end this discussion of almost split sequences with the following global instance of almost split sequences.

PROPOSITION 8 (AUSLANDER-REITEN [10]). *Let X be a connected, Cohen-Macaulay projective curve over an infinite field k . Then an indecomposable Cohen-Macaulay coherent sheaf \mathcal{X} is locally free if and only if there is an almost split sequence $0 \rightarrow \mathcal{F} \rightarrow \mathcal{G} \rightarrow \mathcal{X} \rightarrow 0$ in the category of Cohen-Macaulay coherent sheaves.*

An easy consequence of this result is the following which was first shown independently by Schofield [27] using sheaf theoretic methods and Auslander-Reiten [10] using graded modules.

THEOREM 9. *Let X be a connected nonsingular projective curve over an infinite field. Then the category of coherent sheaves over X has almost split sequences.*

BIBLIOGRAPHY

1. M. Artin, *Maximal orders of global dimension and Krull dimension two*, Invent. Math. (to appear).
2. M. Artin and J.-L. Verdier, *Reflexive modules over rational double points*, Math. Ann. **270** (1985), 71–82.
3. M. Auslander, *Finite type implies isolated singularity* (Proc. Oberwolfach, 1984), Lecture Notes in Math., Vol. 1142, Springer-Verlag, Berlin and New York, 1985, pp. 1–4.
4. —, *Functors and morphisms determined by objects. Applications of morphisms determined by objects* (Proc. Conf. on Representation Theory, Philadelphia, Pa., 1976), Dekker, New York, 1978, pp. 1–327.
5. —, *Isolated singularities and almost split sequences*, Representation Theory. II, Groups and Orders (Ottawa, 1984), Lecture Notes in Math., Vol. 1178, Springer-Verlag, Berlin and New York, 1986, pp. 194–242.
6. —, *Rational singularities and almost split sequences*, Trans. Amer. Math. Soc. **293** (1986), no. 2, 511–532.
7. M. Auslander, M. I. Platzbeck, and I. Reiten, *Coxeter functors without diagrams*, Trans. Amer. Math. Soc. **250** (1979), 1–46.
8. M. Auslander and I. Reiten, *Almost split sequences for Cohen-Macaulay modules* (to appear).
9. —, *Almost split sequences for rational double points*, Trans. Amer. Math. Soc. (to appear).

10. —, *Almost split sequences in dimension 2*, Adv. in Math. (to appear).
11. —, *Grothendieck groups of algebras and orders*, J. Pure Appl. Algebra **39** (1986), 1–51.
12. —, *The Cohen-Macaulay type of Cohen-Macaulay rings* (to appear).
13. —, *Representation theory of artin algebras. III, Almost split sequences*, Comm. Algebra **3** (1975), 239–294.
14. —, *Representation theory of artin algebras. IV, Invariants given by almost split sequences*, Comm. Algebra **5** (1977), 519–554.
15. K. Bongartz and P. Gabriel, *Covering spaces in representation theory*, Invent. Math. **65** (1982), 331–378.
16. S. Brenner and M. C. R. Butler, *Generalizations of the Bernstein-Gelfand-Ponomarev reflection functors*, Representation Theory. II, Lecture Notes in Math., Vol. 832, Springer-Verlag, Berlin and New York, 1980, pp. 103–169.
17. R.-O. Buchweitz, G.-M. Greuel, and F.-O. Schreyer, *Cohen-Macaulay modules on hypersurfaces singularities. II*, Preprint.
18. M. C. R. Butler, *Grothendieck groups and almost split sequences* (Proc. Oberwolfach Conf. on Integral Representations and Applications), Lecture Notes in Math., Vol. 882, Springer-Verlag, Berlin and New York, 1981, pp. 357–368.
19. H. Ésnault, *Reflexive modules on quotient singularities*, J. Reine Angew. Math. (to appear).
20. H. Ésnault and H. Knörrer, *Reflexive modules over rational double points*, Math. Ann. **272** (1985), 545–548.
21. Gonzalez-Springberg and J.-L. Verdier, *Construction geometric de la correspondance de McKay*, Ann. Sci. École Norm. Sup. (4) **16** (1983), 409–449.
22. D. Happel and C. M. Ringel, *Tilted algebras*, Trans. Amer. Math. Soc. **274** (1982), 399–443.
23. J. Herzog, *Ringe mit nur endlich vielen Isomorphieklassen von maximalen, unzerlegbaren Cohen-Macaulay-Moduln*, Math. Ann. **233** (1978), 21–34.
24. H. Knörrer, *Cohen-Macaulay modules on hypersurface singularities. I*, Preprint.
25. J. McKay, *Graphs, singularities and finite groups*, Proc. Sympos. Pure Math., vol. 37, Amer. Math. Soc., Providence, R.I., 1980, pp. 183–186.
26. C. Riedtmann, *Algebren, Darstellungsköcher, Überlagerungen und Zurück*, Comment. Math. Helv. **55** (1980), 199–224.
27. A. Schofield, Private communication.

BRANDEIS UNIVERSITY, WALTHAM, MASSACHUSETTS 02254, USA

О коммутанте абсолютной группы Галуа

Г. В. БЕЛЫЙ

В этом докладе рассматривается один подход к гипотезе о том, что коммутант $G' = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}_{ab})$ группы Галуа $G = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ поля всех алгебраических чисел является свободной проконечной группой. Эта гипотеза была выдвинута И. Р. Шафаревичем.

Пусть H — конечная простая группа. Назовем группу F H -свободной, если для любой точной последовательности конечных групп $1 \rightarrow H^n \hookrightarrow A \twoheadrightarrow B \rightarrow 1$ любой эпиморфизм $F \twoheadrightarrow B$ может быть поднят до эпиморфизма $F \twoheadrightarrow A$. Иными словами, должна существовать пунктирная стрелка:

$$\begin{array}{ccc} & F & \\ \swarrow & & \searrow \\ H^n \hookrightarrow A & \xrightarrow{\quad} & B \end{array}$$

Если доказать, что коммутант G' H -свободен для всех конечных простых групп H , классификация которых недавно получена, будет доказана и эта гипотеза. Ивасава в [I1] доказал, что группа G' H -свободна для всех простых конечных абелевых групп H . Для неабелевых простых конечных групп H доказана

ТЕОРЕМА 1 (Томсон, Матцат). *Если башне накрытий Галуа $X/\mathbb{P}_1/Y$, определенных над полем \mathbb{Q}_{ab} , соответствует точная последовательность групп Галуа $H \hookrightarrow \text{Aut } H \twoheadrightarrow \text{Out } H$, то коммутант G' является H -свободной группой.*

Доказательство см. в [M1].

Пусть F — свободная проконечная группа с двумя образующими x, y . Определим проконечную группу кос B следующим образом:

$$B = \{\sigma \in \text{Aut } F \mid x^\sigma \sim x^\alpha, y^\sigma \sim y^\alpha, (xy)^\sigma \sim (xy)^\alpha, \alpha \in \hat{\mathbb{Z}}^*\},$$

где знак \sim обозначает сопряженность элементов в группе, а $\hat{\mathbb{Z}}^*$ — мультипликативную группу обратимых элементов проконечного пополнения кольца целых чисел; а также ее подгруппы A и A_1 :

$$A = \{\sigma \in B \mid x^\sigma = x^\alpha, y^\sigma = y^{\alpha u}, u \in F'\}, \quad A_1 = \{\sigma \in A \mid \alpha = 1\}.$$

Нетрудно доказать, что централизаторы образующих x, y в группе F являются циклическими группами, изоморфными группе \hat{Z} и, следовательно, $A \cap \text{Int } F = 1$. Это означает, что группа B изоморфна полупрямому произведению $F \rtimes A$. Кроме того, $A/A_1 \cong \hat{Z}^*$.

Пусть $U = P_1 \setminus \{0, 1, \infty\}$ — проективная прямая без трех точек, а K — максимальное неразветвленное вне этих точек расширение поля $\bar{Q}(t)$. Тогда группа Галуа $\text{Gal}(K/\bar{Q}(t))$, совпадающая с проконечным пополнением фундаментальной группы $\pi_1 U(C)$, изоморфна свободной группе F . Оказывается, что группа Галуа $\text{Gal}(K/Q(t))$ изоморфна полупрямому произведению $F \rtimes \pi(G)$, которое определяется некоторым гомоморфизмом $\pi: G \rightarrow A$, причем башне полей $Q(t) \subset \bar{Q}(t) \subset K$ соответствует точная последовательность групп Галуа $F \hookrightarrow F \rtimes \pi(G) \twoheadrightarrow G$. Это легко следует из того обстоятельства, что образующие x, y, xy порождают циклические подгруппы разложения над рациональными точками $0, 1, \infty$, а при действии элементами $\sigma \in \text{Gal}(\bar{Q}(t)/Q(t))$, эти подгруппы инвариантны с точностью до сопряжения. Кроме того, из рассмотрения максимального абелева подрасширения поля K легко получается равенство $\pi(G') = \pi(G) \cap A_1$.

Для эпиморфизма $\varphi: F \twoheadrightarrow H$ в конечную группу справедлива

ТЕОРЕМА 2. Если ядро эпиморфизма φ инвариантно относительно действия группы A_1 и на факторгруппе H это действие индуцирует только внутренние автоморфизмы, то накрытие Галуа X/P_1 , которое соответствует ядру этого эпиморфизма, можно определить над полем Q_{ab} .

Доказательство является легким следствием теории Галуа, см. [B1].

Пусть H — конечная группа без центра с двумя образующими a и b . Рассмотрим множество M , состоящее из пар образующих:

$$M = M(H, a, b) = \{a', b' \in H \mid \langle a', b' \rangle = H, a' \sim a, b' \sim b, a'b' \sim ab\},$$

на котором сопряжениями действует группа H . Если $\#M = \#H$, то есть множество M состоит из одной орбиты, эпиморфизм $\varphi: F \twoheadrightarrow H$, определенный равенствами $\varphi(x) = a$, $\varphi(y) = b$, удовлетворяет условиям теоремы 2. Для проверки этого условия $\#M = \#H$ существует простой критерий: достаточно, чтобы одна из образующих матричной группы H отличалась от скалярной матрицы на матрицу единичного ранга. Этот критерий позволил автору доказать существование расширений $L/Q_{ab}(t)$ с классическими группами Галуа, см. [B1, B2]. Кроме того, с помощью теории представлений конечных групп можно вычислить число $\#M(H, a, b)$, используя таблицы характеров группы H и ее подгрупп. Это позволило Томсону, Матцату и другим авторам доказать существование таких расширений над полем $Q_{ab}(t)$, а в некоторых случаях даже над полем $Q(t)$, для всех спорадических групп, кроме группы J_4 , а также для групп Лиевского типа G_2 , 2G_2 и 3D_4 , см. библиографию в работе [M2]. Необходимо отметить, что наиболее ранней работой, в

которой использовались эти идеи, была работа [S1], где, в частности, построены расширения $L/\mathbb{Q}(t)$ со знакопеременными группами Галуа.

Оказывается, что некоторые из этих расширений удовлетворяют также и условиям теоремы 1. Поэтому группа G' H -свободна, если H — одна из следующих групп:

- (а) знакопеременная группа A_n ($n \neq 6$),
- (б) группа Лиевского типа $A_1, B_n, C_n, D_n, {}^2D_n, G_2$ над простым полем,
- (в) спорадическая группа, кроме J_4 .

Но для многих простых групп H группа $\text{Out } H$ не содержится в группе автоморфизмов проективной прямой P_1 . Поэтому следующим шагом, вероятно, должно быть обобщение этих теорем на случай большей размерности. Теорема 1 легко обобщается, если потребовать, чтобы группа $\text{Out } H$ действовала на проективном пространстве P_n проективными автоморфизмами. Обобщение теоремы 2 остается пока открытой проблемой.

Представление $\pi: G \rightarrow A$ интересно и само по себе. В работе [B1] доказана

ТЕОРЕМА 3. Алгебраическая кривая y может быть определена над полем \mathbb{Q} тогда и только тогда, когда существует непостоянная функция $\varphi \in \mathbb{Q}(y)$, такая, что накрытие $\varphi: y \rightarrow P_1$ разветвлено над тремя точками.

Отсюда легко следует, что представление π является вложением. Очень интересной проблемой является описание его образа. Пусть $\psi: y \hookrightarrow U^n$ — некоторое вложение, определенное над полем \mathbb{Q} . Тогда образ этальной фундаментальной группы $\pi_1(y)$ в группе

$$F^n = \pi_1(U^n)$$

инвариантен с точностью до сопряженности при действии группы G . Не будут ли все инварианты действия группы G на свободной группе F иметь, в этом смысле, геометрическое происхождение?

Пусть F_l — максимальная про- l -факторгруппа группы F , а $F_{l,m}$ — ее максимальная метабелева факторгруппа. Представление π индуцирует соответствующие представления в группах автоморфизмов этих групп. Делинь и Ихара получили очень интересные результаты, относящиеся к этим представлениям, см. [I2].

ЛИТЕРАТУРА

- [B1] Г. В. Белый, *О расширениях Галуа максимального кругового поля*, Изв. АН СССР серия математическая **43** (1979), 267–276. (Math. USSR Izv. **14** (1980), 247–256)
- [B2] G. V. Belyi, *On extensions of the maximal cyclotomic field having a given classical Galois group*, J. Reine Angew. Math. **341** (1983), 147–156.
- [I1] K. Iwasawa, *On solvable extensions of algebraic number fields*, Ann. of Math. **58** (1953), 458–572.

[I2] Y. Ihara, *Profinite braid groups, Galois representations and complex multiplications*, Ann. of Math. **123** (1986), 43–106.

[M1] B. H. Matzat, *Zum Einbettungsproblem der algebraischen Zahlentheorie mit nicht abelschem Kern*, Invent. Math. **80** (1985), 365–374.

[M2] B. H. Matzat, *Über das Umkehrproblem der Galoisschen Theorie*, Karlsruhe, 1985, Preprint.

[S1] K.-Y. Shih, *On the construction of Galois extensions of function fields and number fields*, Math. Ann. **207** (1974), 99–120.

Владимир, СССР

Nilpotent Orbits, Primitive Ideals, and Characteristic Classes (A Survey)

WALTER BORHO

1. Introduction.

1.1. Let G be a semisimple complex Lie group, say linear algebraic, and connected. The present report deals with the following three (a priori) fairly unrelated subjects: (i) The geometry of unipotent conjugacy classes of G , or equivalently, of *nilpotent orbits* in the Lie algebra \mathfrak{g} of G ; (ii) the classification of *primitive ideals* in the universal enveloping algebra $U(\mathfrak{g})$, say with trivial central character for simplicity; and (iii) *characteristic classes* in $H^*(X)$ of certain bundles on the “flag variety” X , that is, the “universal” complete homogeneous space for G . My main point will be to report from recent joint work with J.-L. Brylinski and R. MacPherson [BBM1–3] how (iii) can be used as a tool to get insight into both (i) and (ii) simultaneously, and to understand their relation.

For some time, especially in the seventies and early eighties, these subjects developed more or less independently, each being studied for its own sake, by its own specific methods. Extensive work has been done by many authors, and remarkable theories have been developed on the three subjects, making each of them individually into a highly cultivated area of mathematical research. Since they have been addressed several times at former International Congresses, I feel free here to focus attention on some of the fascinating relations between these subjects. For more background on the individual subjects, the reader may consult, for instance, the Lecture Notes by Steinberg [St], Slodowy [Sl], or Spaltenstein [Sp] on (i), the books by Dixmier [Di] and Jantzen [Ja] on (ii), and say [Hi, Fu] in combination with [BBM1–3] concerning (iii).

1.2. The first hint suggesting that there must be some deep relation between (i) and (ii) became apparent from the fundamental work of T. A. Springer [S] on (i), resp. A. Joseph [J1,2] on (ii): Their results came down to closely relating (i), resp. (ii), to the same kind of objects, namely, *irreducible Weyl group representations*. A careful comparison of Springer’s and Joseph’s correspondences, ultimately extended by D. Barbasch and D. Vogan [BV1,2] to exhaustive explicit case-by-case calculations, confirmed some superficial parallels on one hand,

but also exhibited some intriguing discrepancies on the other hand, so that the real relation remained a mystery for some years.

This situation was considered unsatisfactory by some people, including myself. Strictly speaking, in my case, this challenge dates back already to my 1976 exposé at the Séminaire Bourbaki [B1], where I first suggested relating (i) to (ii) via associated varieties, then reported Jantzen's conjectural partial anticipation of Joseph's theory in case $G = \mathrm{SL}_n$ (see [B1, 2.9, resp. 5.9]), and finally learned from Springer about his new theory; consequently, it was inevitable for me to wonder how these pieces might fit together into a common framework. The solution of this puzzle not only took me some time, but also required some new advanced methods, as well as two good friends to teach me how to use them. Much of my joint work with Brylinski or MacPherson was stimulated by the challenge of this puzzle, and is finally involved in the solution reported here: We use the intersection homology approach to (i), as developed in joint work with MacPherson [BM1,2], and the \mathcal{D} -module approach to (ii), as developed in joint work with Brylinski [BB1,2], and we combine them using equivariant K -theory (on T^*X) as a unifying concept, in order to obtain a common framework for the simultaneous study of (i) and (ii) in terms of (iii), as elaborated in joint work with both [BBM1–3]. Let me also refer at this point to related work of V. Ginsburg [Gi].

1.3. The purpose of this address is two-fold: first, to popularize the “puzzle” mentioned above (§2), and second to state our solution (§§3 and 4). In §2, I tried to illustrate the problem in an intelligible way for the nonexpert, using $G = \mathrm{SL}_n$ as a standard example. Note that this is not *only* a courtesy to the reader not familiar with semisimple Lie group theory, but it also avoids here almost totally those “intriguing discrepancies” mentioned above; this will make our problem (to explain the coincidences between Springer's and Joseph's correspondence) particularly clear and persuasive. On the other hand, those “discrepancies” and their explanations for the other simple Lie groups are precisely what fascinates the real gourmet in Lie theory most of all. So in some sense, the experts may consider it a loss of good taste that I have sacrificed such points radically for the sake of popularity; I hope that they will forgive me.

In §§3 and 4, I tried to formulate an essential part of our results with as little effort as possible. I spent some care in defining important concepts, but no proofs are included here.

2. Review on Springer's and Joseph's Weyl group representations.

2.1. *Basic notation.* We fix a Borel subgroup B in G , and a maximal torus T in B . We denote by $\mathfrak{t} \subset \mathfrak{b} \subset \mathfrak{g}$ the Lie algebras of $T \subset B \subset G$, by $W = N_G(T)/T$ the Weyl group, and by $X = G/B$ the flag variety.

2.2. *Standard example.* To simplify the present review, I shall sometimes restrict the discussion (in the present section only) to $G = \mathrm{SL}_n$ as a standard example. The reader not familiar with general semisimple Lie group theory may anyway prefer to think in terms of this example throughout this paper. Then G

is the group $SL(n, \mathbb{C})$ of complex n -by- n matrices ($n \geq 2$) with determinant 1, and T , resp. B , is the subgroup of its diagonal, resp. upper triangular, matrices. The Lie algebra $\mathfrak{g} = \mathfrak{sl}(n, \mathbb{C})$ consists then of all complex n -by- n matrices of trace 0, with Lie bracket $[x, y] = xy - yx$ the commutator of matrices, and \mathfrak{t} , resp. \mathfrak{b} , consists of all traceless diagonal, resp. upper triangular, matrices. The Weyl group W is the symmetric group, S_n in this case, acting on T and \mathfrak{t} by permuting the n eigenvalues of a diagonal matrix. The flag variety $X = G/B$ may in this case be defined alternatively in the original sense: Its points $F \in X$ may be thought of as real “flags” in \mathbb{C}^n ; that is, F is an ascending chain of complex subspaces $F_1 \subset F_2 \subset \cdots \subset F_n$ of \mathbb{C}^n such that F_i has dimension i .

2.3. Nilpotent orbits. Let \mathcal{N} denote the “nilpotent cone” in \mathfrak{g} , that is, the closed subvariety of all ad-nilpotent elements, which is a cone in \mathfrak{g} . (A cone in a vector space is a subset closed under homotheties.) Under the adjoint action of G on \mathfrak{g} , \mathcal{N} decomposes into a finite number of orbits, called “nilpotent orbits.”

In case $G = SL_n$, these orbits are just the conjugacy classes of nilpotent complex n -by- n matrices. Recall that the set \mathcal{N}/G of nilpotent orbits is here in bijection to the set $\mathcal{P}(n)$ of partitions of n (theory of Jordan normal form), the “parts” being just the sizes of Jordan blocks. For example, the nilpotent orbit \mathcal{O} in \mathfrak{sl}_6 generated by the nilpotent matrix

$$x = \left[\begin{array}{ccc|ccc} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & & & \\ \hline & & & 0 & 1 & \\ & & & & 0 & 1 \\ & & & & & 0 \end{array} \right] \quad (\text{all other entries zero})$$

corresponds to the partition $\lambda = (3, 2, 1)$, which is also denoted

$$\lambda = \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \\ \hline \square & & \\ \hline \end{array}$$

(notation of “Young diagrams”).

2.4. Primitive ideals. By definition, a primitive ideal J in $U(\mathfrak{g})$ is the kernel of some irreducible representation of \mathfrak{g} , or in other words: the annihilator of some simple (left) $U(\mathfrak{g})$ -module L , notation $J = \text{Ann } L$. The center of $U(\mathfrak{g})$ necessarily acts by a character on L , which we assume here to be “trivial,” or equivalently, this means $J \subset \mathfrak{g}U(\mathfrak{g})$. Let \mathcal{X}_0 denote the set of all such primitive ideals. This set is finite. More precisely, one defines a surjection $W \rightarrow \mathcal{X}_0$ as follows: The module L above can always be chosen in a particularly nice way, as a so-called simple highest weight module (theorem of Duflo), and the highest weights in question here come from a single regular W orbit in \mathfrak{t}^* (Harish-Chandra isomorphism), so the modules in question can be suitably indexed by Weyl group elements $w \in W$, and that surjection $W \rightarrow \mathcal{X}_0$ can then be described $w \mapsto \text{Ann } L_w =: J_w$.

2.5. Combinatorial description. In case $G = \mathrm{SL}_n$, the set \mathcal{X}_0 of primitive ideals with trivial central character is in bijection to the set $\mathcal{T}(n)$ of all *tableaux* of size n (theory of Joseph's Goldie rank polynomials, see [Ja]).

Here a “tableau” is the combinatorial object also familiar as a “Young standard tableau” from the representation theory of symmetric groups. An example of a tableau of size $n = 6$ is

$$\tau = \begin{array}{|c|c|c|} \hline 1 & 3 & 5 \\ \hline 2 & 4 & \\ \hline 6 & & \\ \hline \end{array};$$

its “shape” is the partition

$$\lambda = (3, 2, 1) = \begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array},$$

and there are exactly 16 tableaux of this same shape. There is a canonical, easily calculable map $T: W = S_n \rightarrow \mathcal{T}(n)$, producing from each permutation w a tableau $T(w)$ (Robinson-Schensted algorithm). For example, the above tableau τ is produced as $T(w)$ from the permutation $w = 2\,1\,6\,4\,5\,3$, and from 15 other permutations.

Now the bijection between tableaux and primitive ideals, $\mathcal{T}(n) \rightarrow \mathcal{X}_0$, $\tau \mapsto J_\tau$, may be described as follows (Joseph): $J_\tau = \mathrm{Ann} L_w$ for any permutation w of tableau $\tau = T(w)$.

2.6. Associated varieties. There is also a map $\mathcal{X}_0 \rightarrow \mathcal{N}/G$ from primitive ideals to nilpotent orbits. In case $G = \mathrm{SL}_n$, in view of the preceding combinatorial descriptions of the two sets $\mathcal{X}_0, \mathcal{N}/G$, the reader will not find it hard to guess what this map should do in combinatorial terms: It sends the primitive ideal $J = J_\tau$ corresponding to a tableau τ to the nilpotent orbit $\mathcal{O} = \mathcal{O}_\lambda$ corresponding to the shape λ of τ .

But how do we produce directly, and in general, a nilpotent orbit from a primitive ideal J ? For this purpose, we identify the associated graded ring of $U(\mathfrak{g})$ with the symmetric algebra $S(\mathfrak{g})$, and interpret it as a ring of polynomial functions on $\mathfrak{g}^* = \mathfrak{g}$ (using first the Poincaré-Birkhoff-Witt theorem, and second the Killing form). Then we can define the *associated variety* of J as the zero set $\mathcal{V}(\mathrm{gr} J) \subset \mathfrak{g}$ of the associated graded ideal $\mathrm{gr} J \subset S(\mathfrak{g})$. It turns out that this associated variety is *irreducible* [BB1, 2] (see also [J3]) and contained in \mathcal{N} ; hence it contains a unique dense orbit \mathcal{O} . So the desired map $J \mapsto \mathcal{O}$ is given by $\mathcal{V}(\mathrm{gr} J) = \overline{\mathcal{O}}$. Let me comment here that this relation was suggested [B1] *before*, but completely proven [BB1] only *after* Joseph invented his classification theory of primitive ideals (2.9).

2.7. Link to Weyl group representations, illustrated in case $G = \mathrm{SL}_n$. Young diagrams and tableaux, the combinatorial data which we encountered above in the classification of nilpotent orbits, resp. primitive ideals, both have a well-known significance in the (Frobenius) theory of representations of the symmetric

group: There is a bijection, denoted $\lambda \mapsto \rho_\lambda$, $\mathcal{P}(n) \rightarrow W^\wedge$, from diagrams λ of size n to (equivalence classes of) irreducible complex representations of $W = S_n$. Moreover, the tableaux τ of a given shape λ correspond bijectively to a linear *basis* for the representation ρ_λ . In view of this, the results for $G = \mathrm{SL}_n$ reviewed in subsections 2.3 and 2.5, may persuade you to make the following guess:

To a nilpotent orbit \mathcal{O} , there should correspond an irreducible representation $\rho_{\mathcal{O}}$ of the Weyl group, and to the collection of primitive ideals J with associated variety $\overline{\mathcal{O}}$ there should correspond a linear basis of this *same* representation. It turns out that this is actually true, not only in case $G = \mathrm{SL}_n$, but in general, and the main purpose of my report is to explain how such correspondences can be constructed. I shall next introduce the correspondences of Springer and Joseph, which can be made to do this job, although only with some major effort to verify that the two representations are actually the same. An alternative version, where this difficulty disappears, is then stated in Theorems 3.4 and 4.4 below.

2.8. Springer's correspondence. Springer [S] attaches to each nilpotent orbit \mathcal{O} an irreducible representation $\rho_{\mathcal{O}}$ of W as follows: Let $u \in \mathcal{O}$ represent the orbit, and let $X^u \subset X$ be the subvariety of all "flags" respected by u . Then W acts linearly on the homology groups $H_*(X^u)$. (For simplicity, here and below take *complex coefficients* for (co)homology groups, etc.). For $G = \mathrm{SL}_n$, this action is *irreducible* on the homology group $H_{2d}(X^u)$ of highest degree ($2d = 2 \dim X^u$), and this defines $\rho_{\mathcal{O}}$; this is Springer's explicit, *geometrical* realization of the bijection $\mathcal{N}/G \rightarrow W^\wedge$ described only combinatorially in subsection 2.7.

For G arbitrary, one has to replace $H_{2d}(X^u)$ by its invariants under the action of the isotropy group $G_u \subset G$ of u to define Springer's representation $\rho_{\mathcal{O}}$, and one obtains only an *injection* $\mathcal{N}/G \hookrightarrow W^\wedge$ by mapping \mathcal{O} to $\rho_{\mathcal{O}}$.

2.9. Joseph's Goldie rank polynomials. Joseph [J1] attaches to each primitive ideal $J \in \mathcal{X}_0$ a polynomial function p_J on the Cartan subalgebra \mathfrak{t} . His construction proceeds in two steps. The first is to replace J by the whole infinite family $(J_\mu)_{\mu \in Z}$ of "translated" primitive ideals J_μ , with variable central character, depending on a parameter μ which varies in a Zariski dense subset Z of \mathfrak{t}^* ("translation principle" of [BJ]). The second step is to consider the Goldie rank $\mathrm{rk} U(\mathfrak{g})/J_\mu$ as a function of μ , and to prove that this extends to a polynomial function on \mathfrak{t} ; this polynomial, by definition, is p_J . (For the reader not familiar with noncommutative Noetherian ring theory, let me also add the definition of Goldie rank: The quotient ring $U(\mathfrak{g})/J_\mu$ has a complete ring of fractions, which is simple Artinian (Goldie's theorem), hence is a matrix ring over some skew field (Wedderburn-Artin); then $\mathrm{rk} U(\mathfrak{g})/J_\mu$ is the rank of this matrix ring.)

2.10. Joseph's correspondence. Next, Joseph attaches to the primitive ideal $J \in \mathcal{X}_0$ the W submodule generated by p_J in $S(\mathfrak{t}^*)$, and proves that the corresponding W representation, denoted $\sigma(J)$, is irreducible. This defines Joseph's correspondence $\mathcal{X}_0 \rightarrow W^\wedge$, $J \mapsto \sigma(J)$. Moreover, if J' ranges over all primitive ideals such that $\sigma(J') = \sigma(J)$, then the corresponding Goldie rank polynomials provide a *basis* for the representation $\sigma(J)$. The reader will find an excellent exposition of this beautiful theory of Joseph in Jantzen's book [Ja].

2.11. Comparison. The intriguing question, how the correspondences of Springer, resp. Joseph, relate to each other, on which I commented several times before, can be made precise at this point as follows: Does the correspondence from primitive ideals to nilpotent orbits via associated varieties (subsection 2.5) combine with Springer's and Joseph's correspondences to a commutative triangle? Or in other words: Is $\sigma(J)$ equivalent to $\rho_{\mathcal{O}}$, if J corresponds to \mathcal{O} ? In case $G = \mathrm{SL}_n$, the explicit combinatorial descriptions given above prove that this is true. Barbasch and Vogan have verified this as a matter of fact for all cases, by an enormous amount of explicit calculations in [BV1,2]. More conceptual reasons are offered by Hotta and Kashiwara [HK] and below.

3. Characteristic class approach to nilpotent orbits. In these last two sections, I shall now *sketch* the simultaneous, uniform approach to both subjects, (i) nilpotent orbits and (ii) primitive ideals, in terms of (iii) characteristic classes on the flag variety X , as suggested in my recent joint work with Brylinski and MacPherson. For the elaboration of more details, I refer to our original papers [BBM1,2,3].

3.1. Characteristic classes of cone bundles. The concept of a *cone bundle* \mathcal{K} on X generalizes that of a vector bundle: The bundle map $\mathcal{K} \rightarrow X$ is assumed to be a locally trivial fibration of \mathcal{K} by cones (in vector-spaces). To extend the theory of Chern classes of vector-bundles, Fulton and MacPherson [Fu] introduced the notion of *Segre class* $s(\mathcal{K})$ of a cone bundle. It may be characterized by two axioms: (1) for a vector-bundle \mathcal{K} , $s(\mathcal{K}) = c(\mathcal{K})^{-1}$ is the inverse of the total Chern class $c(\mathcal{K})$, and (2) $s(\mathcal{K})$ is functorial under proper push-forward.

Now to define our *characteristic class* $Q(\mathcal{K})$ in $H^*(X)$, for any cone subbundle \mathcal{K} of codimension d in the cotangent bundle T^*X , we multiply its Segre class by the total Chern class of T^*X , and take the lowest (degree $2d$) homogeneous term of the product, or with our chosen notation $Q(\mathcal{K}) := [c(T^*X)s(\mathcal{K})]^{2d}$.

3.2. Springer's resolution of the nilpotent cone. Another key ingredient for our construction is the famous Springer map $\pi: T^*X \rightarrow \mathcal{N}$, which will allow us to pass from nilpotent orbits to the geometry of the flag variety. So let me recall here that this remarkable map has a very easy, elegant definition (as a Kostant-Souriau momentum map [BB1]): The natural action of \mathfrak{g} by vector-fields on X defines a morphism $\mathfrak{g} \times X \rightarrow TX$ into the tangent bundle, and the map π of the cotangent bundle T^*X into $\mathfrak{g}^* = \mathfrak{g}$ (Killing form) is then obtained by dualization and projection. The remarkable point about π is then that its image in \mathfrak{g} is the nilpotent cone \mathcal{N} , and that it resolves the singularities of \mathcal{N} .

3.3. Construction of characteristic classes from a nilpotent orbit. Starting from a nilpotent orbit $\mathcal{O} \subset \mathcal{N}$, we first produce a collection of cone bundles on X by taking the preimage \mathcal{O} under Springer's map π and then decomposing the closure $\overline{\pi^{-1}\mathcal{O}}$ into irreducible components K_1, \dots, K_r . These are in fact cone bundles $K_i \rightarrow X$, called *orbital* for \mathcal{O} . We know that their codimension d in T^*X depends only on \mathcal{O} , or more precisely $2d = \mathrm{codim}_{\mathcal{N}} \mathcal{O}$ (work of Spaltenstein and Steinberg).

Next, we take for these “orbital” cone bundles K_1, \dots, K_r the characteristic classes $Q(K_1), \dots, Q(K_r)$ as defined in subsection 3.1.

- 3.4. THEOREM. (a) *The classes $Q(K_1), \dots, Q(K_r)$ are linearly independent.*
 (b) *They form a W submodule in $H^{2d}(X)$.*
 (c) *This W representation is equivalent to Springer’s ρ_O .*
 (d) *It transforms the basis $Q(K_1), \dots, Q(K_r)$ according to the formulae below.*

The following formulae were first obtained by Hotta in a slightly different context [H1, H2] (see also [J4]).

3.5. *Hotta’s transformation formulae.* For each simple reflection s of W , and for all $i = 1, \dots, r$, we have either

$$sQ(K_i) = -Q(K_i), \quad \text{or else} \quad sQ(K_i) = \sum_{j=1}^r n_{ij}(s)Q(K_j),$$

for some matrix of nonnegative integers $n_{ij}(s)$ with diagonal entries $n_{ii}(s) = 1$. Moreover, there is a geometrical interpretation for these integers, for which I refer to our, resp. Hotta’s, original papers. For example, this says that $n_{ij}(s) = 0$ unless K_j intersects K_i in codimension ≤ 1 .

3.6. *Algebraic construction of our characteristic classes.* For the reader with less inclination for geometrical elegance, but with a preference for abstract algebra, let me also offer here an alternative, more algebraic definition of the classes $Q(K)$ defined geometrically in subsection 3.1. This definition refers to the following ingredients:

- *Borel’s description of $H^*(X)$ in terms of polynomials on the Cartan subalgebra \mathfrak{t} , by a W equivariant isomorphism $\beta: H^*(X) \xrightarrow{\sim} S(\mathfrak{t}^*)^{\natural}$ onto the space of all W harmonic polynomials on \mathfrak{t} ;*
- *the Chern character $\text{ch}: K(X) \rightarrow H^*(X)$, attaching to an (algebraic) vector-bundle on X its total Chern class (an isomorphism here);*
- *the Grothendieck group $K(X)$ of classes of (algebraic) vector-bundles on X , or equivalently the Grothendieck group of the category of all coherent sheaves of \mathcal{O}_X -modules;*
- *the analogous Grothendieck group $K(T^*X)$, and its isomorphism σ^* with $K(X)$ induced by the zero-section $\sigma: X \hookrightarrow T^*X$ [Fu].*

Given these ingredients, we may attach to any closed subvariety K of codimension d in T^*X a cohomology class $Q(K)$ in $H^{2d}(X)$, which we identify with a degree d homogeneous polynomial on \mathfrak{t} via β ($H^{2d}(X) = S^d(\mathfrak{t}^*)^{\natural}$). The construction proceeds as follows: Take the class $[\mathcal{O}_K]$ determined by the structure sheaf of K in $K(T^*X)$, and apply the composition of the maps

$$K(T^*X) \xrightarrow{\sigma^*} K(X) \xrightarrow{\text{ch}} H^*(X) \xrightarrow{\beta} S(\mathfrak{t}^*)^{\natural};$$

finally, take the lowest degree term. So formally, this alternative, algebraic definition of $Q(K)$ reads:

$$Q(K) := [\beta \text{ ch } \sigma^* [\mathcal{O}_K]]^d.$$

For a *cone bundle* \mathcal{K} it can be shown that this definition coincides with that in terms of Segre class as stated in subsection 3.1.

4. Characteristic class of a primitive ideal.

4.1. *Characteristic variety of a primitive ideal.* Starting from a primitive ideal $J \in \mathcal{X}_0$, we first construct a cone-subbundle in T^*X as follows. Later, we shall define the characteristic class of J as the characteristic class of this cone bundle (subsection 4.3).

Consider the quotient $U(\mathfrak{g})/J$ as a left \mathfrak{g} -module M , and take the corresponding sheaf of modules

$$\mathcal{M} := \mathcal{D}_X \otimes_{U(\mathfrak{g})} M$$

over the sheaf \mathcal{D}_X of rings of differential operators on X (Beilinson-Bernstein localization of M on X); then the *characteristic variety*, resp. *cycle*, of \mathcal{M} is a well-defined notion from the general theory of \mathcal{D} -modules, denoted $\text{Ch}(\mathcal{M})$, resp. $\mathcal{C}\mathcal{H}(\mathcal{M})$, here. By definition, $\text{Ch}(\mathcal{M})$ is a closed subvariety in T^*X , and $\mathcal{C}\mathcal{H}(\mathcal{M})$ is a formal integer linear combination

$$\mathcal{C}\mathcal{H}(\mathcal{M}) = \sum_i m_i [\mathcal{V}_i],$$

in which each irreducible component \mathcal{V}_i of the characteristic variety occurs with some well-defined positive multiplicity m_i . Now characteristic varieties are (by construction) fibered by cones, and in the present case, this fibration is locally trivial on X (as a consequence of the stability of J under the adjoint G -action, and the G -homogeneity of X). So $\text{Ch}(\mathcal{M})$, and hence its components \mathcal{V}_i , are actually *cone bundles* over X .

4.2. *Relation to nilpotent orbits.* By my work with Brylinski [BB2], these cone bundles are actually *orbital* for some nilpotent orbit \mathcal{O} . In more detail, by [BB2] Springer's map π maps $\text{Ch}(\mathcal{M})$ onto the associated variety $\mathcal{V}(\text{gr } J)$, which is the closure of a nilpotent orbit, as reported already in subsection 2.6. *In the sequel \mathcal{O} denotes this nilpotent orbit determined by J .* It follows that $\text{Ch}(\mathcal{M})$ is contained in $\pi^{-1}\overline{\mathcal{O}}$, and now some tricky dimension arguments show that each component \mathcal{V}_i is even contained in $\overline{\pi^{-1}\mathcal{O}}$ as one of the irreducible components, hence is one of the orbital cone bundles $\mathcal{K}_1, \dots, \mathcal{K}_r$ (notation of subsection 3.3). Hence we may write our characteristic cycle as well as

$$\mathcal{C}\mathcal{H}(\mathcal{M}) = \sum_{i=1}^r z_i [\mathcal{K}_i],$$

where we admit some of the "multiplicities" z_i to be zero.

4.3. DEFINITION. Now we define the *characteristic class* of J in $H^*(X)$ as the characteristic class of its characteristic variety by:

$$P(J) := Q(\mathcal{C}\mathcal{H}(\mathcal{M})) := \sum_{i=1}^r z_i Q(\mathcal{K}_i).$$

Here $Q(\mathcal{K}_i)$ is the characteristic class of a cone bundle as defined in subsection 3.1 (or 3.6).

4.4. THEOREM. Let $J_1, \dots, J_r \in \mathcal{X}_0$ be the collection of primitive ideals corresponding to a nilpotent orbit \mathcal{O} (that is with associated variety $\overline{\mathcal{O}}$). Then

- (a) the characteristic classes $P(J_1), \dots, P(J_r)$ are linearly independent.
- (b) They span a W submodule in $H^{2d}(X)$. (Note $2d = \text{codim}_{\mathcal{N}} \mathcal{O}$.)
- (c) This W representation is equivalent to Springer's $\rho_{\mathcal{O}}$.
- (d) The class $P(J_i)$ is proportional to Joseph's Goldie rank polynomial, that is, $\beta P(J_i) = \gamma p_{J_i}$ for some scalar $0 \neq \gamma \in \mathbb{Q}$.

REFERENCES

- [BV1] D. Barbasch and D. Vogan, *Primitive ideals and orbital integrals in complex classical groups*, Math. Ann. **259** (1982), 153–199.
- [BV2] ———, *Primitive ideals and orbital integrals in complex exceptional groups*, J. Algebra **80** (1983), 350–382.
- [B1] W. Borho, *Recent advances in enveloping algebras of semisimple Lie algebras*, Séminaire Bourbaki, exp. 489, Lecture Notes in Math., vol. 677, Springer-Verlag, 1978, pp. 1–18.
- [BJ] W. Borho and J. C. Jantzen, *Über primitive Ideale in der Einhüllenden einer halbeinfachen Lie-Algebra*, Invent. Math. **39** (1977), 1–53.
- [BB1] W. Borho and J. L. Brylinski, *Differential operators on homogeneous spaces. I*, Invent. Math. **69** (1982), 437–476.
- [BB2] ———, *Differential operators on homogeneous spaces, III*, Invent. Math. **80** (1985), 1–68.
- [BBM1] W. Borho, J. L. Brylinski, and R. MacPherson, *A note on primitive ideals and characteristic classes*, Geometry Today, Progress in Math., vol. 60, Birkhäuser, 1985, pp. 11–20.
- [BBM2] ———, *Springer's Weyl group representations through characteristic classes of cone bundles*, Preprint, Inst. Hautes Études Sci. M/85/70, Dec. 1985.
- [BBM3] ———, *Equivariant K-theory approach to nilpotent orbits*, Preprint, Inst. Hautes Études Sci. M/86/13, March 1986.
- [BM1] W. Borho and R. MacPherson, *Représentations des groupes de Weyl et homologie d'intersection pour les variétés nilpotentes*, C. R. Acad. Sci. Paris Sér. A **292** (1981), 707–710.
- [BM2] ———, *Partial resolutions of nilpotent varieties*, Analyse et Topologie sur les Espaces Singuliers, Astérisque, no. 101, Soc. Math. France, Paris, 1983, pp. 23–74.
- [Di] J. Dixmier, *Algèbres enveloppantes*, Gauthier-Villars, Paris, 1974.
- [Fu] W. Fulton, *Intersection theory*, Springer-Verlag, 1984.
- [Gi] V. Ginsburg, *\mathfrak{g} -modules, Springer's representations and bivariant Chern classes*, Adv. in Math. **59** (1986).
- [H] H. Hiller, *Geometry of Coxeter groups*, Res. Notes in Math., vol. 54, Pitman, Boston-London-Melbourne, 1982.
- [H1] R. Hotta, *On Joseph's construction of Weyl group representations*, Tôhoku Math. J. **36** (1984), 49–74.
- [H2] ———, *On Springer's representations*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. **28** (1982), 836–876.
- [HK] R. Hotta and M. Kashiwara, *The invariant holonomic system on a semisimple Lie algebra*, Invent. Math. **75** (1984), 327–358.
- [Ja] J. C. Jantzen, *Einhüllende Algebren halbeinfacher Lie-Algebren*, Springer-Verlag, 1983.
- [J1] A. Joseph, *Goldie rank in the enveloping algebra of a semisimple Lie algebra, I, II*, J. Algebra **65** (1980), 269–306.
- [J2] ———, *Kostant's problem, Goldie rank, and the Gelfand-Kirillov conjecture*, Invent. Math. **56** (1980), 191–213.
- [J3] ———, *On the associated variety of a primitive ideal*, J. Algebra **93** (1985), 509–523.

- [J4] —, *On the variety of a highest weight module*, J. Algebra **88** (1984), 238–278.
- [S1] P. Slodowy, *Simple singularities and simple algebraic groups*, Lecture Notes in Math., vol. 815, Springer-Verlag, 1980.
- [Sp] N. Spaltenstein, *Classes unipotentes et sous-groupes de Borel*, Lecture Notes in Math., vol. 946, Springer-Verlag, 1982.
- [St] R. Steinberg, *Conjugacy classes in algebraic groups*, Lecture Notes in Math., vol. 366, Springer-Verlag, 1974.
- [S] T. A. Springer, *Trigonometric sums, Green functions of finite groups, and representations of Weyl groups*, Invent. Math. **36** (1976), 173–207.

BUGH WUPPERTAL, GAUßSTRASSE 20, 5600 WUPPERTAL 1, FEDERAL REPUBLIC OF GERMANY

Théorie locale des blocs d'un groupe fini

MICHEL BROUÉ

1. Les p -blocs d'un groupe fini et la matrice de décomposition généralisée.

A. *Les blocs: définition.* Soit p un nombre premier et soit G un groupe fini. On désigne par K une extension finie du corps des nombres p -adiques \mathbb{Q}_p , d'anneau des entiers \mathcal{O} et de corps résiduel k . On suppose dans tout ce qui suit que K est "assez gros" pour G , i.e., que \mathcal{O} contient les racines de l'unité d'ordre l'exposant du groupe (ce qui implique que les algèbres de groupe KG et kG sont déployées).

L'algèbre KG est semi-simple et se décompose en un produit direct d'algèbres de matrices sur K indexées par l'ensemble des classes d'isomorphisme de représentations irréductibles de G sur K (cf. [Se]), donc par l'ensemble $\text{Irr}(G, K)$ des caractères irréductibles de G sur K :

$$KG \xrightarrow{\sim} \prod M_K(\chi) \quad (\chi \in \text{Irr}(G, K)).$$

L'algèbre kG n'est pas semi-simple dès que p divise l'ordre de G . On note JkG son radical de Jacobson. Sa "semi-simplifiée" kG/JkG se décompose en un produit direct d'algèbres de matrices sur k indexées par l'ensemble des classes d'isomorphisme de représentations irréductibles de G sur k , i.e., par l'ensemble $\text{IBr}(G, \mathcal{O})$ des caractères de Brauer (ou caractères modulaires) irréductibles de G sur \mathcal{O} (cf. [Se]):

$$kG/JkG \xrightarrow{\sim} \prod M_k(\varphi) \quad (\varphi \in \text{IBr}(G, \mathcal{O})).$$

L'algèbre $\mathcal{O}G$ se décompose en un produit direct d'algèbres indécomposables appelées les blocs de $\mathcal{O}G$ (ou blocs de G sur \mathcal{O}). Cette décomposition est fournie par la décomposition de 1 en idempotents orthogonaux primitifs dans le centre $Z\mathcal{O}G$ de $\mathcal{O}G$: si b est un idempotent primitif de $Z\mathcal{O}G$, le bloc de $\mathcal{O}G$ correspondant est $\mathcal{O}Gb$. Puisque l'anneau \mathcal{O} est complet, si B est un bloc de $\mathcal{O}G$, la k -algèbre $k \otimes_{\mathcal{O}} B$ est indécomposable; elle est appelée bloc de kG (ou bloc de G sur k). Dans toute la suite, on appellera "bloc de G " soit un bloc de $\mathcal{O}G$, soit un bloc de kG . Enfin pour tout bloc B de $\mathcal{O}G$, on note $\text{Mod}(B)$ la catégorie des

B -modules à gauche, libres et de rang fini sur \mathcal{O} .

$$\begin{array}{ccccc} & & KG & \xrightarrow{\sim} & \prod K \otimes_{\mathcal{O}} B \\ & \nearrow & & & \\ \mathcal{O}G & \xrightarrow{\sim} & \prod B & & \\ & \searrow & & & \\ & & kG/JkG & \xrightarrow{\sim} & \prod B/JB \end{array}$$

Soit B un bloc de $\mathcal{O}G$. On note $\text{Irr}(G, B)$ l'ensemble des caractères irréductibles χ de G sur K tels que la projection $B \rightarrow M_K(\chi)$ ne soit pas nulle. On a alors

$$K \otimes_{\mathcal{O}} B \xrightarrow{\sim} \prod M_K(\chi) \quad (\chi \in \text{Irr}(G, B)).$$

On note de même $\text{IBr}(G, B)$ l'ensemble des caractères modulaires irréductibles φ tels que la projection $B \rightarrow M_k(\varphi)$ ne soit pas nulle, et on a

$$B/JB \xrightarrow{\sim} \prod M_k(\varphi) \quad (\varphi \in \text{IBr}(G, B)).$$

Les blocs de G définissent donc des partitions de l'ensemble des caractères irréductibles de G sur K et de l'ensemble des caractères modulaires irréductibles de G . Pour plus de détails (et pour d'autres définitions des partitions ainsi définies), on peut se reporter à [Fe].

On appelle bloc principal de G et on note $B_0(G)$ le bloc tel que $\text{Irr}(G, B_0(G))$ contienne le caractère trivial de G sur K . Il est facile de vérifier que

$$\text{IBr}(G, B_0(G))$$

contient alors le caractère modulaire de la représentation triviale de G sur k .

B. *La matrice de décomposition généralisée.* Tout au long des quelques trente années de travail que R. Brauer a consacrées à la théorie des représentations modulaires des groupes finis (cf. [Bra]), il a concentré son attention sur les relations entre les caractères irréductibles, la "structure p -locale" (cf. ci-dessous, §2) et les caractères modulaires, en particulier en étudiant l'objet mathématique qui exprime ces relations: la matrice de décomposition généralisée.

La "table généralisée des caractères modulaires" est par définition la matrice $(a_{(x, \varphi)}^{(g)})$ définie de la manière suivante

g parcourt un système de représentants des classes de conjugaison d'éléments de G ,

x parcourt un système de représentants des classes de conjugaison des p -éléments de G , et φ parcourt l'ensemble des caractères modulaires irréductibles du centralisateur $C_G(x)$ de x dans G ,

$a_{(x, \varphi)}^{(g)}$ vaut 0 si la p -composante de g n'est pas conjuguée à x , et vaut $\varphi(x')$ si x' et x sont respectivement la p' -composante et la p -composante de g .

La matrice de décomposition généralisée, notée ici $\text{Dec}(G, p)$, est la matrice de passage entre la table des caractères ordinaires et la table généralisée des caractères modulaires définie ci-dessus. Elle est donc égale à la matrice $(d_{\chi}^{(x, \varphi)})$ où les nombres $d_{\chi}^{(x, \varphi)}$ sont définis par

$$\chi(xx') = \sum_{\varphi \in \text{IBr}(C_G(x), \mathcal{O})} d_{\chi}^{(x, \varphi)} \varphi(x')$$

pour tout $\chi \in \text{Irr}(G, K)$, x p -élément de G , x' p' -élément de $C_G(x)$. Il est facile de voir que $d_\chi^{(x, \varphi)} \in \mathbb{Z}[\zeta_p]$ où ζ_p désigne une racine de l'unité d'ordre la p -partie de l'exposant de G .

Le morphisme de Brauer (cf. ci-dessous, §2) permet de répartir les classes de conjugaison de paires (x, φ) entre les différents blocs de G (* Si B est un bloc de G , on note $(x, \varphi) \in B$ si φ appartient à l'ensemble $\text{IBr}(C_G(x), E)$ où E est un bloc de $C_G(x)$ tel que $(\langle x \rangle, E)$ soit une B -souspaire de G *). Le "Second Théorème Principal de Brauer" établit que la matrice de décomposition généralisée est décomposée en blocs diagonaux de matrices carrées indexées par les blocs de G :

$$\text{Dec}(G, p) = \left[\begin{array}{ccc} \text{diagonal blocks} \end{array} \right] \rightarrow \text{Dec}(G, B) = (d_\chi^{(x, \varphi)}), \quad (x, \varphi) \in B, \chi \in \text{Irr}(G, B).$$

La partie (non carrée en général) de la matrice correspondant aux indices (x, φ) avec $x = 1$ est la matrice de décomposition usuelle (cf. [Se]).

2. Structure locale sur un bloc. Soit P un p -sous-groupe de G . On note $(kG)^P$ l'ensemble des points de kG fixes sous l'action de P par conjugaison et on note $C_G(P)$ le centralisateur de P dans G . Le résultat suivant, essentiellement dû à Brauer (cf. [Al-Br] pour une démonstration de la forme présentée ici) est à la base de l'étude locale des blocs.

(2.1) THÉORÈME-DÉFINITION. Soit $\text{Br}_P: (kG)^P \rightarrow kC_G(P)$ l'application induite par l'application linéaire de kG sur $kC_G(P)$ qui envoie $g \in G$ sur g ou 0 selon que g appartient ou non à $C_G(P)$. L'application Br_P est un morphisme d'algèbres, appelé morphisme de Brauer.

(2.2) DÉFINITION [Al-Br]. On appelle souspaire de G toute paire (P, E) où P est un p -sous-groupe de G et E est un bloc de $C_G(P)$.

On a une notion naturelle d'inclusion entre souspaires [Al-Br], qui peut être définie de la manière suivant.

(2.3) THÉORÈME [Br-Pu]. Soient P et Q deux p -sous-groupes de G tels que $Q \subset P$. Soit E un bloc de $kC_G(P)$. Il existe un (unique) bloc F de $kC_G(Q)$ possédant la propriété suivante: soit i un idempotent primitif de $kC_G(P)$ tel que $i \in E$. Si \hat{i} est un idempotent primitif de l'algèbre $(kG)^P$ d'image i par le morphisme Br_P , on a $\text{Br}_Q(\hat{i}) \in F$.

$$\begin{array}{ccc} (kG)^P & \hookrightarrow & (kG)^Q \\ \downarrow \text{Br}_P & & \downarrow \text{Br}_Q \\ kC_G(P) & \hookrightarrow & kC_G(Q) \end{array}$$

On dit alors que la souspaire (Q, F) est contenue dans la souspaire (P, E) et on note $(Q, F) \subset (P, E)$.

De plus, on dit que (Q, F) est normale dans (P, E) si Q est normal dans P et si $(Q, F) \subset (P, E)$; cette dernière condition est réalisée si et seulement si $\text{Br}_P(F) \subset E$. Il est clair d'après (2.3) que la relation d'inclusion est une relation d'ordre, et que si $(Q, F) \subset (P, E)$, alors (Q, F) est sous-normal dans

(P, E) . Nous allons voir que l'inclusion des souspaires généralise l'inclusion des p -sous-groupes de G .

Soit B un bloc de G . On dit qu'une souspaire (P, E) est une B -paire si $(\{1\}, B)$ est contenue dans (P, E) . Dans le cas particulier où B est le bloc principal de G , on a le résultat suivant, sorte de reformulation du "Troisième Théorème Principal de Brauer."

(2.4) THÉORÈME [Al-Br]. *L'application $P \mapsto (P, B_0(C_G(P)))$ est une bijection de l'ensemble des p -sous-groupes de G sur l'ensemble des $B_0(G)$ -sous-paires de G qui respecte les relations d'inclusion.*

Suivant [Pu1], on appelle catégorie de Frobenius de G pour p et on note $\mathfrak{Fr}(G, p)$ la catégorie dont les objets sont les p -sous-groupes de G , et telle que les morphismes de Q dans P soient les homomorphismes de Q dans P induits par les automorphismes intérieurs de G . De manière analogue (et, en fait, plus générale: cf. (2.5) ci-dessous), si B est un bloc de G , on appelle catégorie de Brauer de G pour B et on note $\mathfrak{Br}(G, B)$ la catégorie dont les objets sont les B -souspaires de G , et telle que les morphismes de (Q, F) dans (P, E) soient les homomorphismes de Q dans P induits par les automorphismes intérieurs de G qui envoient (Q, F) sur une souspaire contenue dans (P, E) .

De même que l'étude p -locale d'un groupe fini est l'étude de $\mathfrak{Fr}(G, p)$ [Pu1], l'étude locale d'un bloc B est l'étude de sa catégorie de Brauer $\mathfrak{Br}(G, B)$. Le théorème (2.4) ci-dessus peut se ré-interpréter de la manière suivante:

(2.5) *L'application $P \mapsto (P, B_0(C_G(P)))$ identifie*
 $\mathfrak{Fr}(G, p)$ *et* $\mathfrak{Br}(G, B_0(G)).$

Ainsi, l'étude locale des blocs apparaît comme une généralisation de l'étude p -locale des groupes finis. De fait, nombre de résultats connus se généralisent, tels les théorèmes de Sylow, qui peuvent être vus comme une reformulation du "Premier Théorème Principal de Brauer."

(2.6) THÉORÈME ([Al-Br] "THÉORÈMES DE SYLOW"). (1) *Soit B un bloc de G . Le groupe G opère transitivement par conjugaison sur l'ensemble des B -souspaires maximales de G .*

(1) *(caractérisation locale de la maximalité) Soit (P, E) une souspaire de G et soit $N_G(P, E)$ son stabilisateur dans G . Les assertions suivantes sont équivalentes:*

- (i) (P, E) *est une souspaire maximale de G ,*
- (ii) (P, E) *est une souspaire maximale de $N_G(P, E)$.*

(2.7) DÉFINITION. Si (P, E) est une B -souspaire maximale de G , le groupe P est appelé un groupe de défaut de B .

Les groupes de défaut de B forment (d'après (2.6)(1)) une classe de conjugaison de p -sous-groupes de G .

Un bloc B de groupe de défaut le groupe trivial $\{1\}$ se comporte "comme si p ne divisait pas l'ordre de G "; plus précisément, le bloc B a pour groupe

de défaut $\{1\}$ si et seulement si l'une des conditions équivalentes suivantes est réalisée:

- (i) L'algèbre B est une algèbre de matrices sur \mathcal{O} ;
- (ii) la catégorie de modules $\text{Mod}(B)$ est équivalente à la catégorie $\text{Mod}(\mathcal{O})$ des modules libres et de rang fini sur \mathcal{O} .

Cette situation "triviale" donne une indication sur le cas général: la structure (à équivalence près) de la catégorie de Brauer $\mathfrak{Br}(G, B)$ influe sur la structure (à équivalence près) de la catégorie de modules $\text{Mod}(B)$ (cf. ci-dessous (2.12)).

Le théorème de fusion d'Alperin (1967) (précisé par Goldschmidt [Go] et Puig [Pu0]) se généralise aussi en remplaçant p -sous-groupes par B -sous-paires.

(2.8) DÉFINITIONS. (1) Soit (D, E) une sous-paire maximale. Une famille \mathcal{C} de sous-paires contenues dans (D, E) est dite famille de conjugaison pour (D, E) si elle possède la propriété suivante: pour toute paire (P, E_P) et tout élément g de G tels que (P, E_P) et $(P, E_P)^g$ soient contenus dans (D, E) , il existe $(Q_1, F_1), \dots, (Q_n, F_n)$ dans \mathcal{C} et $g_i \in N_G(Q_i, F_i)$ ($i = 1, 2, \dots, n$), $z \in C_G(P)$ tels que

$$g = zg_1g_2 \cdots g_n, \quad (P, E_P) \subset (Q_1, F_1), \\ (P, E_P)^{g_1 \cdots g_{i-1}} \subset (Q_i, F_i) \quad \text{pour } i = 2, \dots, n.$$

(2) Une sous-paire (Q, F) est dite essentielle si

(a) elle est maximale comme sous-paire du groupe $QC_G(Q)$,

(b) le groupe $N_G(Q, F)/QC_G(Q)$ contient un sous-groupe propre p -fortement plongé.

(On dit que le sous-groupe H de G est p -fortement plongé dans G si p divise l'ordre de H mais ne divise l'ordre d'aucun des groupes $H \cap H^g$ pour $g \in G - H$.)

(2.9) THÉORÈME. Une famille \mathcal{C} de sous-paires contenues dans une sous-paire maximale (D, E) est une famille de conjugaison pour (D, E) si et seulement si \mathcal{C} contient un représentant de chaque classe de conjugaison de sous-paires essentielles.

Le théorème de Burnside concernant le cas où les p -sous-groupes de Sylow sont abéliens se généralise aisément à la catégorie de Brauer.

(2.10) PROPOSITION. Soit (D, E) une B -sous-paire maximale de G . On suppose D abélien. Alors l'application $P \mapsto (P, E_P)$ où $(P, E_P) \subset (D, E)$ définit une équivalence de catégories entre $\mathfrak{F}(D \rtimes (N_G(D, E)/C_G(D)), p)$ et $\mathfrak{Br}(G, B)$.

Dans le cas où B est le bloc principal (i.e., $\mathfrak{Br}(G, B) = \mathfrak{F}(G, p)$), on n'en déduit rien de plus sur la structure du groupe G . Mais l'étude de la catégorie $\text{Mod}(B)$ des B -modules permet d'aller un peu plus loin, y compris dans le cas où $B = B_0(G)$. Si par exemple D est cyclique, la catégorie de modules $\text{Mod}(B)$ "ressemble" à la catégorie des modules sur $(D \rtimes N_G(D, E)/C_G(D))$. Si elle ne lui est pas équivalente en général, elle a néanmoins le même nombre d'indécomposables, d'irréductibles, et est bien décrite par l'arbre de Brauer (cf. [Da, Fe]). De plus, les matrices de décomposition généralisées $\text{Dec}(G, B)$

et $\text{Dec}(D \rtimes (N_G(D, E)/C_G(D)), p)$ ont même taille et sont égales "aux signes près":

$$\text{Dec}(G, B) = \begin{bmatrix} \pm 1 & & 0 \\ & \pm 1 & \\ 0 & & \pm 1 \end{bmatrix} \cdot \text{Dec}(D \rtimes N_G(D, E)/C_G(D), p).$$

Ce dernier fait a également été vérifié dans tous les exemples connus de blocs tels que D soit abélien et $D \rtimes N_G(D)/C_G(D)$ soit un groupe de Frobenius de noyau D . Ces hypothèses impliquent en tout cas (Puig) l'équivalence stable des catégories $\text{Mod}(B)$ et $\text{Mod}(D \rtimes N_G(D, E)/C_G(D))$. Plus généralement, il est conjecturé (Alperin) que $\text{Dec}(G, B)$ et $\text{Dec}(D \rtimes N_G(D, E)/C_G(D), p)$ ont la même taille lorsque D est abélien (cf. (3.3) ci-dessous).

Un des résultats les plus remarquables de la théorie locale des groupes finis est dû à Frobenius (1907). Avant de l'énoncer, remarquons que si $O_{p'}(G)$ désigne le plus grand sous-groupe normal de G d'ordre premier à p , la surjection canonique de G sur $G/O_{p'}(G)$ définit une équivalence de catégories entre $\mathfrak{Ft}(G, p)$ et $\mathfrak{Ft}(G/O_{p'}(G), p)$. Le théorème de Frobenius établit que si $\mathfrak{Ft}(G, p)$ est équivalente à la catégorie d'un p -groupe, alors G lui-même (modulo $O_{p'}(G)$) est un p -groupe. Plus précisément:

(2.11) THÉORÈME (FROBENIUS). *Si $N_G(P)/C_G(P)$ est un p -groupe pour tout p -sous-groupe P de G , alors $G/O_{p'}(G)$ est un p -groupe.*

On peut voir ce théorème comme un résultat concernant le bloc principal, de la manière suivante. Soit D un p -sous-groupe de Sylow de G , et soit $b_0(G)$ l'idempotent primitif de $Z\mathcal{O}G$ correspondant au bloc principal de G . Alors $G/O_{p'}(G)$ est un p -groupe si et seulement si l'application $x \mapsto xb_0(G)$ est un isomorphisme entre $\mathcal{O}D$ et le bloc principal $\mathcal{O}Gb_0(G)$.

Ce résultat a pu être généralisé au cas d'un bloc quelconque [Br-Pu, Pu3], sous la forme suivante:

(2.12) THÉORÈME. *Soit B un bloc de G , de groupe de défaut D . Supposons que $N_G(P, E)/C_G(P)$ soit un p -groupe pour toute B -sous-paire (P, E) . Alors $\text{Mod}(B)$ est équivalente à $\text{Mod}(\mathcal{O}D)$. De plus, les matrices $\text{Dec}(G, B)$ et $\text{Dec}(D, p)$ sont "égales aux signes près":*

$$\text{Dec}(G, B) = \begin{bmatrix} \pm 1 & & 0 \\ & \pm 1 & \\ 0 & & \pm 1 \end{bmatrix} \cdot \text{Dec}(D, p).$$

Il est à noter que la structure de D lui-même n'intervient pas dans l'énoncé précédent, et donc que ce théorème met en évidence l'importance cruciale de la structure de $\mathfrak{Bt}(G, B)$ pour la structure de $\text{Mod}(B)$. Cela étant, la structure du groupe D influe sur celle de la catégorie $\mathfrak{Bt}(G, B)$, comme le montre le cas où D est cyclique (ou, plus simplement, le cas où $D = \{1\}$). De plus, les résultats obtenus concernent bien la structure de la catégorie de modules $\text{Mod}(B)$ et non celle de groupe G : il existe des blocs de groupes simples satisfaisant aux hypothèses du théorème (2.12).

Ce résultat met aussi en évidence les limites de l'étude locale: il ramène le problème à celui de l'étude des représentations d'un p -groupe. Ce dernier problème requiert alors d'autres méthodes, en particulier homologiques, comme le montrent par exemple les travaux récents de J. Carlson.

D'autres résultats généraux de la théorie locale des groupes finis fournissent des conjectures naturelles pour le cas d'un bloc quelconque. Il en est ainsi du Z^* - p -théorème, dû à Glauberman [G1] pour $p = 2$, et conséquence de la classification des groupes finis simples pour p impair.

(2.13) THÉORÈME. *Soit P un p -sous-groupe de G tel que l'inclusion de $C_G(P)$ dans G induise une équivalence de catégories entre $\mathfrak{F}r(C_G(P), p)$ et $\mathfrak{F}r(G, p)$. Alors $G = C_G(P)O_p(G)$.*

Un cas particulier (dû à Brauer et Suzuki) du théorème précédent est obtenu en considérant un groupe dont un 2-sous-groupe de Sylow D est quaternionien. Il est facile de vérifier que, si P est l'unique sous-groupe d'ordre 2 de D , les hypothèses de (2.13) sont satisfaites. Cela suggère la question suivante:

(2.14) QUESTION. Soit $p = 2$, et soit B un bloc de G dont les groupes de défaut sont quaternioniens. Soit P l'unique sous-groupe d'ordre 2 d'un groupe de défaut, et soit E un bloc de $C_G(P)$ tel que (P, E) soit une B -souspaire de G . Les catégories de modules $\text{Mod}(B)$ et $\text{Mod}(E)$ sont-elles équivalentes?

3. Quelques problèmes ouverts.

A. *Les algèbres de source.* La théorie locale des blocs et celle des "vertices" et sources des modules indécomposables (due à J. A. Green) ont été englobées par Puig dans le langage des groupes pointés [Pu2]. Dans ce cadre, la notion d'algèbre de source d'une G -algèbre intérieure indécomposable joue un rôle crucial (cf. [Pu3], par exemple). Nous nous contentons de la définir ici pour les blocs.

Soit B un bloc de G et soit (D, E) une B -souspaire maximale. Il est immédiat de vérifier que le bloc E de $C_G(D)$ vérifie les hypothèses du théorème "de Frobenius" (2.13), et par conséquent $\text{Mod}(E)$ est équivalente à la catégorie des modules sur un p -groupe; en particulier, il n'y a (à isomorphisme près) qu'un seul E -module projectif indécomposable.

Considérons alors la situation décrite dans l'énoncé du théorème (2.3). Soit i un idempotent primitif de $kC_G(D)$ tel que $i \in k \otimes_O E$ (ainsi, le $(k \otimes_O E)$ -module $(k \otimes_O E)i$ est projectif indécomposable), et soit \hat{i} un idempotent primitif de $(kG)^D$ d'image i par Br_D . Soit enfin j un idempotent primitif de $(OG)^D$ d'image \hat{i} dans $(kG)^D$.

On note $\text{Sce}(B)$ et on appelle algèbre de source de B l'algèbre jBj munie de l'homomorphisme de groupes $D \rightarrow \text{Sce}(B)^\times$ défini par $x \mapsto xj$. Puisque i est unique (à conjugaison près par $(kC_G(D))^\times$), le couple $(D, \text{Sce}(B))$ est ainsi défini de manière unique à G -conjugaison près.

(3.1) THÉORÈME (PUIG). (1) *Les catégories de modules*
 $\text{Mod}(B)$ *et* $\text{Mod}(\text{Sce}(B))$

sont équivalentes.

(2) La connaissance de $\text{Sce}(B)$ permet de reconstituer $\mathfrak{Br}(G, B)$ à équivalence près.

(3) La connaissance de $\text{Sce}(B)$ permet de reconstituer $\text{Dec}(G, B)$.

La notion d'algèbre de source peut être un outil commode vers une classification en théorie des blocs, grâce au théorème précédent et à la conjecture suivante.

(3.2) CONJECTURE (PUIG). Pour tout p -groupe D , il n'existe qu'un nombre fini de couples (S, π) (à isomorphisme près) où S est une \mathcal{O} -algèbre et π un homomorphisme $D \rightarrow S^\times$ qui soient les algèbres de source de blocs de groupes de défaut D .

Cette conjecture est vérifiée, grâce à la classification des groupes finis simples, dans le cas où on se restreint aux blocs des groupes p -résolubles.

A l'appui de cette conjecture, voici un exemple d'une famille infinie de blocs ayant tous même groupe de défaut et même algèbre de source.

Soit r un nombre premier tel que p divise $r - 1$, et soit (a_n) une suite infinie d'entiers non divisibles par p . Soit enfin m un entier tel que $m < p$. On pose $G_n = \text{GL}_m(r^{a_n})$, et on note B_n le sous-groupe de G_n formé des matrices triangulaires supérieures; on pose

$$i_n = \frac{1}{|O_{p'}(B_n)|} \sum_{x \in O_{p'}(B_n)} x.$$

Soit D un p -sous-groupe du Sylow de G_n , isomorphe au produit direct de m copies du p -sous-groupe de Sylow de $(\mathbf{Z}/r\mathbf{Z})^\times$. Il est facile de vérifier que i_n est un idempotent primitif de $(\mathcal{O}G_n)^D$ et que $i_n \mathcal{O}G_n i_n$ est algèbre de source du bloc principal de G_n . D'autre part, $i_n \mathcal{O}G_n i_n$ est "constante" et isomorphe à $\mathcal{O}(D \rtimes N_{G_1}(D)/C_{G_1}(D))$.

B. *La conjecture d'Alperin*. Comme indiqué au paragraphe précédent, l'expérience suggère que certains des résultats connus sur les blocs à groupe de défaut cyclique peuvent s'étendre à une situation plus générale. D'autre part, il serait important, en vue d'une simplification de la démonstration du théorème de classification des groupes finis simples, de posséder une démonstration *a priori* du Z^* - p -théorème (cf. (2.14) ci-dessus) dans le cas où p est impair.

Un pas important dans ces deux directions serait fourni par la vérification de la conjecture d'Alperin, qui affirme que l'on peut "calculer localement" la taille de la matrice de décomposition $\text{Dec}(G, B)$. Plus précisément:

(3.3) CONJECTURE (ALPERIN). Le nombre de k -représentations irréductibles d'un bloc B de G est égal à

$$\sum_{(P, E)} n_0(P, E),$$

où $n_0(P, E)$ est le nombre de classes d'isomorphismes de $kN_G(P, E)$ -modules irréductibles sur lesquels l'élément unité de E agit comme l'identité et qui sont des $k(N_G(P, E)/P)$ -modules projectifs, et où (P, E) parcourt un système de représentants des classes de conjugaison (sous G) de B -souspaires de G .

Cette conjecture a été vérifiée dans le cas des groupes p -résolubles par Dade [Da2] et par Okuyama et Wajima [Ok], dans le cas des groupes de type de Lie en caractéristique p par Cabanes, dans le cas des groupes symétriques par Alperin et dans celui des groupes linéaires en caractéristique différente de p par Fong.

C. *Les blocs des groupes de type de Lie en caractéristique non naturelle.* Le travail de Fong et Srinivasan [Fo-Sr] a mis en évidence les relations remarquables qui existent entre les outils de la théorie des blocs et l'induction généralisée introduite par Deligne et Lusztig. Ces relations ont permis [Bro] de déterminer la structure de $\mathfrak{Br}(G, B)$ pour $G = \mathrm{GL}_n(q)$ ou $\mathrm{U}_n(q^2)$, et B un l -bloc de G avec $l \nmid p$. Il reste à généraliser ces résultats aux autres "groupes finis réductifs," et à étudier les exemples nombreux qu'ils fournissent et les généralisations qu'ils suggèrent.

BIBLIOGRAPHIE

- [Al-Br] J. L. Alperin et M. Broué, *Local methods in block theory*, Ann. of Math. **110** (1979), 143–157.
- [Be] D. Benson, *Modular representation theory: new trends and methods*, Lecture Notes in Math., vol. 1081, Springer-Verlag, Berlin-New York, 1984.
- [Bra] R. Brauer, *Collected papers*. I, II, III, M.I.T. Press, Cambridge, Mass., 1980.
- [Bro] M. Broué, *Les l -blocs des groupes $\mathrm{GL}(n, q)$ et $\mathrm{U}(n, q^2)$ et leurs structures locales*, Sémin. Bourbaki, 37^e année (1984/85), n°640, Astérisque **133–134** (1986), 159–188.
- [Br-Pu] M. Broué et Ll. Puig, *A Frobenius theorem for blocks*, Invent. Math. **56** (1981), 117–128.
- [Da1] E. C. Dade, *Blocks with cyclic defect groups*, Ann. of Math. (2) **84** (1966), 20–48.
- [Da2] —, *A correspondance of characters*, Proc. Sympos. Pure Math., vol. 37, Amer. Math. Soc., Providence, R.I., 1980, pp. 401–404.
- [Fo-Sr] P. Fong et B. Srinivasan, *The blocks of finite general linear and unitary groups*, Invent. Math. **69** (1982), 109–153.
- [Fe] W. Feit *The representation theory of finite groups*, North-Holland, Amsterdam, 1982.
- [Go] D. M. Goldschmidt, *A conjugation family for finite groups*, J. Algebra **16** (1970), 138–142.
- [Gl] G. Glauberman, *Central elements in core-free groups*, J. Algebra **4** (1966), 403–420.
- [Gr] J. A. Green, *Some remarks on defect groups*, Math. Z. **107** (1968), 133–150.
- [Ok] T. Okuyama, *Module correspondence in finite groups*, Hokkaido Math. J. **10** (1981), 299–318.
- [Pu0] Ll. Puig, *Structure locale dans les groupes finis*, Bull. Soc. Math. France **47** (1976).
- [Pu1] —, *La classification des groupes finis simples: bref aperçu et quelques conséquences internes*, Sémin. Bourbaki, 34^e année (1981/82), n°584, Astérisque **92–93** (1984), 101–128.
- [Pu2] —, *Pointed groups and construction of characters*, Math. Z. **176** (1981), 265–292.
- [Pu3] —, *The source algebra of a nilpotent block*, Preprint, 1982.
- [Pu4] —, *Local fusions in block source algebras*, J. Algebra (à paraître).
- [Se] J.-P. Serre, *Représentations linéaires des groupes finis*, 3^e édition, Hermann, Paris, 1978.

Equivariant Embeddings of Homogeneous Spaces

CORRADO DE CONCINI

0. Introduction. Given an algebraic group G over an algebraically closed field k , which, for most of this paper, will be assumed of characteristic 0, and a closed subgroup $H \subset G$, an embedding of the homogeneous space G/H is a pair (X, i) , where X is a G -variety and i a G -equivariant embedding $i: G/H \rightarrow X$ with $i(G/H)$ an open dense subset of X .

For a general homogeneous space the structure of embeddings is largely unknown and mysterious except in some cases (for a beautiful beginning see [LV]). Here we shall review some results on embeddings of a special type of homogeneous space that is called spherical. By definition a homogeneous space G/H is called spherical [BLV, B1] if G is a connected reductive group and a Borel subgroup $B \subset G$ has a dense orbit in G/H .

This class of homogeneous spaces, although quite “small,” (for a classification when H is reductive see [B1]) contains many interesting examples such as tori, homogeneous spaces G/H with H a maximal unipotent subgroup of G (see [P1] where a classification of embeddings in this case is given), or H the subgroup of elements fixed by an automorphism of order two of G (in this case one calls G/H a symmetric variety).

The paper will be organized in a way respectful of the “history” of the subject. So we shall start by recalling a few facts about the, by now, well-known theory of torus embeddings [D1, KMS, Da, O], which plays a special role in the theory of embeddings of spherical homogeneous spaces since, as we shall see below, one can in a certain sense reduce to the case of a torus. Then we shall review the case of symmetric varieties [DP1, DP2], which is on the one hand motivated by some examples arising from classical enumerative geometry and on the other motivates, at least for the author, further developments. We shall then briefly explain some recent results [BLV] in the general case of a spherical space. Finally we shall define an “intersection ring” for spherical homogeneous spaces, which gives sound foundations for Schubert calculus on such spaces.

We finish this introduction by giving some of the notation needed below.

Given two embeddings $(X_1, i_1), (X_2, i_2)$ of a homogeneous space G/H , a morphism $\phi: (X_1, i_1) \rightarrow (X_2, i_2)$ is a morphism of G -varieties $\phi: X_1 \rightarrow X_2$ with the

property that the diagram

$$\begin{array}{ccc} G/H & \xrightarrow{i_1} & X_1 \\ \downarrow \text{id} & & \downarrow \phi \\ G/H & \xrightarrow{i_2} & X_2 \end{array}$$

commutes.

It is clear that a morphism between two embeddings, if it exists, is unique, so we can give a structure of partially ordered set to the set of (isomorphism classes of) embeddings of G/H by setting $(X_1, i_1) \geq (X_2, i_2)$ if a morphism $\phi: (X_1, i_1) \rightarrow (X_2, i_2)$ exists. Usually in what follows, when considering an embedding (X, i) , we shall omit the injection and denote it only by X .

1. Review of the theory of torus embeddings. As already remarked, the theory of torus embeddings is by now well known and here we shall only give a brief review of it. In this section the field k can be taken of arbitrary characteristic. Let $T \cong G_m^n$, let G_m the multiplicative group of k , be a torus, and let $X^*(T)$ be its group of characters. Consider the group $X_*(T) = \text{Hom}(X^*(T), \mathbf{Z})$ of one-parameter subgroups of T together with the real vector space $V = \text{Hom}(X^*(T), \mathbf{R}) \supset X_*(T)$. The lattice $X_*(T)$ induces a rational structure on V .

Given an affine embedding A of T , we associate to it a rational pointed cone σ_A of V as follows: Let $S \subset X^*(T)$ be the subset consisting of characters χ with the property that there exists a T -proper vector in the coordinate ring of A transforming according to χ ; we define $\sigma_A = \{v \in V \mid \langle v, \chi \rangle \geq 0, \forall \chi \in S\} \subset V$.

It is then easily seen that, if we assume A to be normal, the cone σ_A determines A and the map $A \rightarrow \sigma_A$ gives a bijection between normal affine embeddings of T and rational pointed cones.

Let us now consider a general normal T -embedding Z . By [S] we can cover Z with open affine T -stable subsets. Given an affine open T -stable subset $A \subseteq Z$, we have seen how to associate a cone σ_A to A . Thus to Z we can associate a (finite) collection $\Sigma_Z = \{\sigma_A\}$, A varying over the (finite) set of open affine T -stable subsets of Z .

The collection Σ_Z satisfies the following two properties:

(a) If $\sigma \in \Sigma_Z$ and τ is a face of σ , then $\tau \in \Sigma_Z$.

(b) If $\sigma, \tau \in \Sigma_Z$, $\sigma \cap \tau$ is a face of both σ and τ .

We now define a rational partial polyhedral decomposition (briefly r.p.p.d.) of V to be a collection Σ of rational pointed cones in V satisfying properties (a) and (b) above.

PROPOSITION. *The mapping $Z \rightarrow \Sigma_Z$ gives a bijection between normal T -embeddings and r.p.p.d.'s.*

In fact one can read many of the geometric properties of Z from the associated r.p.p.d. We list a few here.

(1) Given an open affine T -stable subset $A \subseteq Z$, there is a unique T -orbit $O_A \subseteq A$ which is relatively closed in A . Vice versa, any T -orbit in Z is relatively

closed in a unique affine T -stable open subset of Z . In this way one gets a bijection between T -orbits in Z and cones in Σ_Z . Furthermore, given two orbits O_1 and O_2 of Z and letting σ_1 and σ_2 be the corresponding faces, $O_1 \subseteq \overline{O_2}$ if and only if σ_2 is a face of σ_1 .

(2) Z is complete if and only if V is the union of the cones in Σ_Z .

(3) Z is smooth if and only if for each cone $\sigma \in \Sigma_Z$ there exist vectors $v_1, v_2, \dots, v_r \in X_*(T)$ such that $\sigma = \{v \in V | v = \sum a_i v_i, a_i \geq 0\}$ and v_1, v_2, \dots, v_r can be completed to a basis of $X_*(T)$.

(4) $Z_1 \geq Z_2$ if and only if every cone in Σ_{Z_1} is contained in a cone in Σ_{Z_2} . (In this case we set $\Sigma_{Z_1} \geq \Sigma_{Z_2}$.) Furthermore the morphism $\phi: Z_1 \rightarrow Z_2$ is proper if and only if each cone in Σ_{Z_2} is the union of cones in Σ_{Z_1} .

2. Symmetric varieties. In this section G will be a semisimple group of adjoint type, $\sigma: G \rightarrow G$ an order two automorphism of G , $H = G^\sigma = \{g \in G | \sigma(g) = g\}$.

In this case G/H , which is, as one can easily deduce from the Iwasawa decomposition, a spherical homogeneous space, has a canonical projective embedding X . X has various definitions [DP1, Sp]. Here we shall give the quickest and most natural one, which was inspired by [D2], where some special cases were treated, warning the reader that it is by no means the easiest to work with.

Let $\mathfrak{g} = \text{Lie } G$, $\mathfrak{h} = \text{Lie } H$, $\dim \mathfrak{h} = s$. We can consider \mathfrak{h} as a point in the Grassman variety $\mathcal{G}_s(\mathfrak{g})$ of s -dimensional subspaces of \mathfrak{g} . The adjoint action of G on \mathfrak{g} induces an action on $\mathcal{G}_s(\mathfrak{g})$. We define X as the closure of the orbit $G \circ \mathfrak{h} \subseteq \mathcal{G}_s(\mathfrak{g})$. Since H equals its normalizer in G , we then get that X is indeed a projective embedding of G/H .

The study of X depends on the existence of a very nice open set in X . In order to explain this we need to recall a few facts. A torus $S \subseteq G$ is called split if, $\forall s \in S$, $\sigma(s) = s^{-1}$. Split tori exist [V1]. Fix a maximal split torus T_1 and a maximal torus $T \subseteq G$ containing T_1 . One knows that T is stable under σ ; thus σ induces an involution on the character group $X^*(T)$ which preserves the root system \mathcal{R} . We define a map $\mu: X^*(T) \rightarrow X^*(T)$ by $\mu(\chi)(t) = \chi(t)\chi(\sigma(t^{-1}))$. Restricting characters to T_1 we see that we can identify $\mu(X^*(T))$ with $X^*(T')$, where $T' = T_1/T_1 \cap H \cong T/T \cap H$. (Notice that $T_1 \cap H$ is the subgroup of T_1 of elements of order two.)

Furthermore $\mathcal{R}' = \mu(\mathcal{R}) - \{0\} \subset X^*(T')$ is a (possibly nonreduced) root system. Let W_r be its Weyl group. Consider the vector space

$$V = \text{Hom}(X^*(T'), \mathbf{R}) \supset X_*(T').$$

The root system \mathcal{R}' induces a rational polyhedral decomposition Σ of V whose cones are the Weyl chambers and their faces. Associated to Σ , we then get a smooth and complete embedding Z of T' . Notice that Σ is W_r -stable, i.e., if $\sigma \in \Sigma$, $w \in W_r$, $w \circ \sigma \in \Sigma$. This implies that W_r acts on Z , extending the natural action of W_r on T' . If we choose a fundamental Weyl chamber σ , and hence an ordering on the root system \mathcal{R}' , we can associate to σ an affine T' -stable

open subset $Z^{(o)}$ of Z which is isomorphic to an affine space, and a parabolic subgroup $P \subset G$, namely, the parabolic associated to the set of roots $\alpha \in \mathcal{R}$ such that $\mu(\alpha) \geq 0$. We denote by P^u the unipotent radical of P .

Consider now the T -orbit $T \circ [H] \subseteq G/H$. It is easy to see that $T \circ [H]$ is isomorphic to T' , so for an embedding Y of G/H , the closure Z_Y of $T \circ [H]$ in Y is an embedding of T' . Furthermore we have [He] that $W_r \cong N_H(T_1)/C_H(T_1)$, where $N_H(T_1)$ (resp. $C_H(T_1)$) denotes the normalizer (resp. centralizer) of T_1 in H , so that W_r acts on Z_Y , extending the natural action of W_r on T (we shall call such an embedding an embedding with W_r -action). Going back to X we have

LEMMA [DP1]. *There exists an affine P -stable open set $A \subseteq X$ meeting every G -orbit in X such that*

- (1) $A \cap Z_X \cong Z^{(o)}$.
- (2) *The map $P^u \times Z^{(o)} \cong P^u \times (A \cap Z_X) \rightarrow A$ induced by the P^u action on A is an isomorphism; in particular, A is isomorphic to an affine space.*
- (3) *For every G -orbit O in X there exists a unique T -orbit O' in $Z^{(o)}$; such $O \cap A$ is the isomorphic image of $P^u \times O'$ under the isomorphism given in (2).*

From this lemma it is then easy to deduce many properties of X .

THEOREM [DP1]. (1) *The closure of every G -orbit in X is smooth (in particular, X is smooth).*

(2) *$X - G \circ \mathfrak{h}$ is a divisor D with normal crossings and each irreducible component of D is the closure of a G -orbit.*

(3) *Let D_1, D_2, \dots, D_m , $m = \dim T_1$, be the irreducible components of D . Then the closure of any G -orbit O is the transversal intersection of the D_i 's containing O .*

(4) $\bigcap_{i=1}^r D_i \cong G/P$ *is the unique closed orbit in X .*

We now define \mathcal{E} to be the partially ordered set of embeddings of G/H which are greater than or equal to X and \mathcal{T} to be the set of embeddings with W_r -action of T' which are greater than or equal to Z . The following proposition, which is an easy consequence of the above theorem and its proof, tells us that a lot of information about the geometry of an embedding $Y \in \mathcal{E}$ can be deduced by looking at the torus embedding Z_Y .

PROPOSITION [DP2]. (1) $Z_X = Z$. *In particular, for any $Y \in \mathcal{E}$, $Z_Y \in \mathcal{T}$.*

(2) *The map $\mathcal{J}: \mathcal{E} \rightarrow \mathcal{T}$, defined by $\mathcal{J}(Y) = Z_Y$, is an isomorphism of partially ordered sets.*

(3) *An embedding $Y \in \mathcal{E}$ is normal (resp. complete, smooth, projective) if and only if Z_Y is normal (resp. complete, smooth, projective). In particular, by the theory of torus embeddings we get a bijection between the normal embeddings in \mathcal{T} and the set of W_r -stable r.p.p.d. of V which are greater than or equal to Σ .*

(4) *Let $Y \in \mathcal{E}$. The intersection of Z_Y with any G -orbit $O \subseteq Y$ is proper, nonempty, and is the disjoint union of T' -orbits which are permuted transitively by W_r . Given two G -orbits O_1 and O_2 we have that $O_1 \subseteq \overline{O_2}$ if and only if $Z_Y \cap O_1 \subseteq \overline{Z_Y \cap O_2}$.*

Notice that in the case when $G = H \times H$ and $\sigma(h_1, h_2) = (h_2, h_1)$, $G/H \cong H$, considered as a $H \times H$ -space under the action given by left and right multiplication, so in this case the variety X is a canonical compactification of the group H , which we shall denote by \overline{H} . (In this case the theory of embeddings is closely related to the algebraic monoids of [Pu, Re].)

It turns out that we can embed the canonical compactification X of any symmetric variety G/H , in \overline{G} as follows: we consider the map $\mu: G \rightarrow G$ defined by $\mu(g) = g\sigma(g^{-1})$; then $\mu(G)$ is a closed subvariety of G isomorphic to G/H [R], and its closure in \overline{G} is exactly X .

At this point a few final remarks are in order.

(a) When $X = \overline{\text{PGL}(n)}$, X is called the variety of complete projectivities of \mathbf{P}^{n-1} . When $G = \text{PGL}(n)$ and σ is orthogonal involution, G/H is the variety of nondegenerate quadrics in \mathbf{P}^{n-1} and X is called the variety of complete quadrics. Both these varieties have been introduced and studied by classical geometers (see [T, Se] and for more modern treatments [V, La]) in relation to enumerative problems such as the computation of characteristic numbers for quadrics. An algorithm for computation of “characteristic numbers” in the case of a general symmetric variety has been given in [DP1] and implemented on a computer [DGT].

(b) A variety analogous to the variety X has been introduced in [OS] in the context of real symmetric spaces. The construction given in [OS] has been translated into the algebraic contest to give an alternative construction of X in [Sp].

(c) In [V2] Vust has announced a complete classification of normal embeddings of G/H in terms of “colored r.p.d.’s” using the theory developed in [LV].

(d) In the “case of a group” one can prove the results stated in this section under no assumption on the characteristic (see [St] where other properties of \overline{H} are proved in a characteristic-free contest). In general the theory works under the assumption that the base field has characteristic not 2 using [U], but the proofs are substantially different from those given in [DP1].

3. Spherical spaces. In this section we shall review the results on [BLV] which show how to generalize at least some of the results of the preceding section to general spherical spaces.

The results in [BLV] about spherical homogeneous spaces are a consequence of a very general result which we are going to state. Let G be a reductive connected group, Z a normal G -variety, $z \in Z$ a point with the property that the orbit $G \circ z$ is projective, so that the stabilizer of $G_z \subseteq G$ is parabolic. Let P be a parabolic subgroup opposite to G_z and $L = G_z \cap P$. L is of course a Levi factor of P . Let P^u denote the unipotent radical of P . Then we have:

THEOREM [BLV]. *There exists a locally closed affine variety $W \subseteq Z$ such that*

- (a) $z \in W$ and W is L -stable;
- (b) $P^u \circ W$ is an open set in Z ;

(c) *the map $P^u \times W \rightarrow P^u \circ W$, given by the action of P^u on Z , is an isomorphism.*

This result is proved by reducing the analysis, by a result of [S], to the case in which G acts linearly on a vector space N , and we take $Z = \mathbf{P}(N)$, with G -action induced by the linear action of G on N .

In order to clarify the relation with the results of §2, we notice that, using the notations of §2, if we take Z to be the canonical compactification X of a symmetric variety, z to be the center of the open Schubert cell in the closed orbit of X , we can take $W = A \cap Z_X \cong Z^{(o)}$.

We can now briefly explain which results about spherical homogeneous space are deduced in [BLV] from this theorem.

Let G/H be a spherical homogeneous space, and let $B \subseteq G$ be a Borel subgroup in G , with the property that $BH \subseteq G$ is open in G . One can easily check that $G - BH$ is a subvariety which is pure of codimension 1 in G . We denote by $\mathcal{P}_{++} \subset k[G]$, the subset of regular functions whose set of zeros is exactly (set-theoretically) $G - BH$ and have value 1 on the identity of G . We now want to define a parabolic subgroup $P \supseteq B$ of G and a torus in P which will play, in this case, a role analogous to the role played for symmetric varieties by the parabolic subgroup associated to a fundamental Weyl chamber for \mathcal{R}' and by a maximal split torus.

P is easily defined. We let $P = \{g \in G | gBH = BH\}$. It is then clear that $P \supseteq B$. As for the torus, its definition is more subtle. We choose $f \in \mathcal{P}_{++}$ and consider the differential $df(e)$, e the identity of G . $df(e) \in \mathfrak{g}^*$, $\mathfrak{g} = \text{Lie } G$. G acts on \mathfrak{g}^* by the coadjoint action and we let $L^f \subseteq G$ be the stabilizer of $df(e)$ in G . We now define our torus C^f to be the connected center of L^f . The fact that C^f is a torus follows since one shows [BLV] that L^f is a Levi subgroup of P and hence reductive.

By a repeated application of the above theorem one then gets:

THEOREM [BLV]. (1) $L^f \cap H = P \cap H$ and this group is reductive.

(2) *The derived subgroup of L^f is contained in H .*

(3) *Given any G -variety Z and a point $z \in Z$ fixed by H , the subset $P^u \circ \overline{(C^f \circ z)}$ contains a nonempty open set of every G -orbit in $\overline{G \circ z}$.*

Notice that (3) tells us exactly that the torus embedding $\overline{C^f \circ z}$ meets every orbit in $\overline{G \circ z}$; in particular, this applies for embeddings of G/H , that is, when $G_z = H$.

Let Y be a G/H embedding and Z_Y be the embedding of $C^f/C^f \cap H$ which is the closure of $C^f \circ [H]$ in Y . From the above theorem we see immediately that Y contains finitely many G -orbits (in effect each G -orbit of Y meets Z_Y in a nonempty union of C^f -orbits and there are finitely many such orbits), and each G -orbit is itself a spherical homogeneous space, so by [B2] we get in fact that Y has finitely many B -orbits.

From these results one can hope to describe the G/H -embeddings, Y , in terms of the torus embeddings Z_Y . From the methods used in the proof of the theorem

quite a bit can be said (see also [BL]). On the other hand, it is not at all clear (for some examples see [P1, P2]) which torus embeddings appear as Z_Y 's for some G/H embedding Y . (In the case of a symmetric variety the Weyl group W_r played an important role but this seems to have no analogue in general.) The situation is complicated by the fact that the choice of $f \in \mathcal{P}_{++}$ is arbitrary and there is not, up to now, a clear "best choice" (one is proposed in [BLV]).

4. Schubert calculus. Let us consider a general homogeneous space $G/H = M$, with G connected and let $n = \dim M$. By Kleiman's transversality theorem [K] we have that, given two irreducible subvarieties $V_1, V_2 \subseteq M$, the intersection $V_1 \cap g \circ V_2$ is a proper intersection with multiplicity one in each component for g belonging to a nonempty open set U of G . When V_1 and V_2 have complementary codimensions in M we can choose U in such a way that for $g \in U$ the intersection $V_1 \cap V_2$ consists of finitely many simple points whose number, which we shall denote by (V_1, V_2) , is independent of the choice of g in U .

It is clear that we can extend the pairing $(,)$ to arbitrary cycles by linearity. Thus denoting by $Z^r(M)$ the group of codimension r cycles in M we get a bilinear pairing,

$$(,): Z^r(M) \times Z^{n-r}(M) \rightarrow \mathbb{Z}.$$

We then set $\mathcal{B}^r(M) = \{a \in Z^r(M) | (a, b) = 0, \forall b \in Z^{n-r}(M)\}$ and set $C^r(M) = Z^r(M)/\mathcal{B}^r(M)$ and $C^*(M) = \bigoplus_r C^r(M)$, so that $C^*(M)$ is a graded abelian group together with a bilinear pairing,

$$(,): C^r(M) \times C^{n-r}(M) \rightarrow \mathbb{Z},$$

induced by the pairing on Z^* , which is nondegenerate in the sense that if $a \in C^r(M)$ is such that $(a, b) = 0 \forall b \in C^{n-r}(M)$, then $a = 0$. The question we want to address is the following:

Does there exist a ring structure on $C^*(M)$ having the property that given $V_1, V_2 \subseteq M$ as above and denoting by $[V_1]$ and $[V_2]$ their classes in $C^*(M)$, $[V_1][V_2] = [V_1 \cap gV_2]$ with g generic?

This is in general not possible (for an easy example see [DP2]). However if we are in the case of a spherical homogeneous space it is possible to define this structure.

Consider the partially ordered set \mathcal{S} of smooth projective embeddings of G/H . If $Y \in \mathcal{S}$, then since Y contains finitely many G -orbits, any maximal torus $T \subseteq G$ has finitely many fixpoints in Y . Hence using the theory of [BB] we have that Y has a "cellular decomposition" by locally closed affine cells. From this it is easy to verify that, denoting by $A^*(Y)$ (resp. $H^*(Y)$) the Chow ring (resp. the cohomology ring) of Y , Y has no odd cohomology and the natural map $A^*(Y) \rightarrow H^{2*}(Y)$ is an isomorphism. Recall that if $Y_1 \leq Y_2$ we get a unique morphism $\pi: Y_1 \rightarrow Y_2$ and hence a canonical homomorphism $A^*(Y_2) \rightarrow A^*(Y_1)$.

We can then consider the direct limit,

$$\mathcal{Y}^*(G/H) = \varinjlim A^*(Y) \cong \varinjlim H^{2*}(Y), \quad Y \in \mathcal{S}.$$

Any irreducible subvariety $V \subseteq G/H$ has a well-defined class in $\mathcal{H}^*(G/H)$. Indeed one shows [DP2] that there exists $Y \in \mathcal{S}$ such that $\bar{V} \subseteq Y$ has proper intersection with every G -orbit in Y , and if we consider the class of \bar{V} in $A^*(Y)$ its image (V) in $\mathcal{H}^*(G/H)$ is independent of the choice of Y with the above properties.

We then have

THEOREM [DP2]. *There is a natural isomorphism*

$$\Phi: \mathcal{H}^*(G/H) \rightarrow C^*(G/H),$$

such that $\Phi([V]) = [V]$ for every irreducible subvariety $V \subseteq G/H$. In particular, $C^(G/H)$ inherits a ring structure from $\mathcal{H}^*(G/H)$ which has the properties required above.*

In [DP2] the result is proved for symmetric varieties; however one can see, using the results stated in §2, that the proofs go on verbatim in the spherical case.

The computation of the ring \mathcal{H}^* is a complicated matter even in the case of a torus where at least the computation of the cohomology ring of smooth projective embeddings is completely known [Da]. In general even the knowledge of the cohomology ring of a smooth projective embedding of spherical homogeneous space is very limited. The only examples known to the author apart from torus embeddings are the case of the “compactifications of a group” [DP3], and of complete quadrics [DGMP] where the rational cohomology is known (for conics see also [CX]).

Betti numbers have been computed for the canonical embedding of a symmetric variety [St1, DS].

REFERENCES

- [BB] A. Bialynicki-Birula, *Some theorem on actions of algebraic groups*, Ann. of Math. (2) **98** (1973), 480–497.
- [B1] M. Brion, *Classification des espaces homogènes sphériques*, Bull. Soc. Math. France (to appear).
- [B2] —, *Quelques propriétés des espaces homogènes sphériques*, Manuscripta Math. **55** (1986), 191–198.
- [BL] M. Brion and D. Luna, *Sur la structure locale des variétés sphériques*, Preprint.
- [BLV] M. Brion, D. Luna and Th. Vust, *Espaces homogènes sphériques*, Invent. Math. **84** (1986), 617–632.
- [CX] E. Casas and S. Xambó, *The enumerative theory of conics after Halphen*, Lecture Notes in Math., vol. 1196, Springer-Verlag, Berlin-New York.
- [Da] V. I. Danilov, *The geometry of toric varieties*, Russian Math. Surveys **33** (1978), 97–154.
- [DGT] C. De Concini, P. Gianni and C. Traverso, *Computation of new Schubert tables for quadrics and projectivities*, Adv. Stud. Pure Math. **6** (1985), 515–523.
- [DGMP] C. De Concini, M. Goreski, R. McPherson, and C. Procesi, *On the geometry of quadrics and their degenerations*, Comment. Math. Helv. (to appear).
- [DP1] C. De Concini and C. Procesi, *Complete symmetric varieties*, Lecture Notes in Math., vol. 996, Springer-Verlag, Berlin-New York, 1983, pp. 1–44.
- [DP2] —, *Complete symmetric varieties. II*, Adv. Stud. Pure Math. **6** (1985), 481–513.

- [DP3] —, *Cohomology compactifications of a semisimple adjoint group*, Duke Math. J. **53**, no. 3, 585–596.
- [DS] C. De Concini and T. A. Springer, *Betti numbers of complete symmetric varieties*, Progr. Math., vol. 60, Birkhäuser, 1985, pp. 87–119.
- [D1] M. Demazure, *Sous-groupes algébriques de rang maximum du groupe de Cremona*, Ann. Sci. École Norm. Sup. **3** (1971), 507–588.
- [D2] —, *Limites de groupes orthogonaux ou symplectiques*, Preprint (1980).
- [He] S. Helgason, *Differential geometry, Lie groups and symmetric spaces*, Academic Press, 1978.
- [KMS] G. Kempf, F. Knudsen, D. Mumford, and B. Saint-Donat, *Toroidal embeddings*, Lecture Notes in Math., vol. 339, Springer-Verlag, Berlin-New York, 1974.
- [K] S. Kleiman, *The transversality of a general translate*, Compositio Math. **28** (1974), 287–297.
- [L] D. Laksov, *Complete linear maps. I: The geometry*, Preprint.
- [LV] D. Luna and Th. Vust, *Plongements d'espaces homogènes*, Comment. Math. Helv. **58** (1983), 186–245.
- [O] T. Oda, *Lectures on torus embeddings and applications*, Tata Inst. Fund. Res. Lecture Notes XI, Tata Inst. Fund. Res., Bombay, 1978.
- [OS] T. Oshima and J. Sekiguchi, *Eigenspaces of invariant differential operators in affine symmetric spaces*, Invent. Math. **57** (1980), 1–81.
- [P1] F. Pauer, *Normale einbettungen von G/U* , Math. Ann. **257** (1981), 371–396.
- [P2] —, *Plongements normaux de l'espace homogène $SL(3)/Sl(2)$* , C. R. du 108e Con. nat. Soc. sav. Grenoble (1983).
- [Pu] M. Putcha, *Linear algebraic semigroups*, Semigroup Forum **22** (1981), 287–309.
- [Re] L. E. Renner, *Classification of semisimple algebraic monoids*, Trans. Amer. Math. Soc. **292** (1985), 193–223.
- [R] R. W. Richardson, *Orbits invariants and representations associated to involutions of reductive groups*, Invent. Math. **66** (1982), 287–312.
- [Se] J. G. Semple, *The variety whose points represent complete collineations of S_r on S'_r* , Univ. Roma Ist. Naz. Alta Mat. Rend. Mat. e Appl. **10** (1951), 201–208.
- [Sp] T. A. Springer, *Algebraic groups with involutions*, Preprint.
- [St1] E. Strickland, *Schubert type cells for complete quadrics*, Adv. in Math. (to appear).
- [St2] —, *A vanishing theorem for group compactifications in arbitrary characteristic*, Preprint.
- [S] H. Sumihiro, *Equivariant completion*, J. Math. Kyoto Univ. **14** (1974), 1–28.
- [T] J. A. Tyrell, *Complete quadrics and collineations in S_n* , Mathematika **3** (1956), 69–79.
- [U] To. Uzawa, *On the equivariant completions of algebraic symmetric spaces*, Alg. Top. Theories (1985), 569–577.
- [V] I. Vainsencher, *Schubert calculus for complete quadrics*, Enumerative geometry and classical algebraic geometry (Nice, 1981), Progr. Math., vol. 24, Birkhäuser, 1982, pp. 199–235.
- [V1] Th. Vust, *Opération des groupes réductifs dans un type de cônes presque homogènes*, Bull. Soc. Math. France **102** (1974), 317–333.
- [V2] —, *Communication in Basel*, 1982.

UNIVERSITÀ DI ROMA, TOR VERGATA, ITALY

Darstellungen endlichdimensionaler Algebren

PETER GABRIEL

Wir betrachten hier *endlichdimensionale* assoziative Algebren A mit Eins über einem algebraisch abgeschlossenen Körper k . Moduln über A sind stets als Rechtsmoduln zu verstehen. Die Algebra A heisst *darstellungsendlich*, wenn sie nur endlich viele Isomorphieklassen (direkt) unzerlegbarer Moduln besitzt. Im kommutativen Bereich sind solche Algebren sehr dünn gesät: zugelassen sind nur endliche Produkte mit Faktoren der Gestalt $k[T]/T^n$. Es wimmelt aber das nicht-kommutative Gefilde von darstellungsendlichen Algebren. Ihnen widmen wir diesen Bericht. Vom Standpunkt der Darstellungstheorie bilden sie die erste Klasse, die in Angriff zu nehmen ist, so man Klassifikationsprobleme liebt.

Wir bringen Kunde von einem darstellungsendlichen Paradies. Leider bewohnen Algebren der Alltagspraxis meist die Hölle, wovon uns nur Randzonen bekannt sind mit ihren noch zahmen Wesen. Ihnen galt die Aufmerksamkeit Ringels in seinem Warschauer Bericht [43]. Wir verweisen auf dort angekündigte, inzwischen erschienene Notizen [44], deren Form uns nicht leicht fällt.

1. Die kombinatorischen Invarianten darstellungsender Algebren.

1.1. Eines der zentralen Ergebnisse unseres Berichts besagt, dass es in jeder Dimension nur endlich viele Isomorphieklassen darstellungsender Algebren gibt. Dies befiehlt uns, zur Beschreibung dieser Algebren Ausschau zu halten nach diskreten Parametern, nach kombinatorischen Invarianten, die wir gleich zu Beginn einführen.

Zunächst dürfen wir dank der Morita-Äquivalenz [36] annehmen, dass A eine *Basialgebra* ist, d.h. dass die Restklassenalgebra nach dem Radikal die Gestalt $k^n = k \times \cdots \times k$ hat. Ist dann $1 = \sum_{e \in E} e$ eine möglichst feine Zerlegung der Eins von A mit $e^2 = e$ und $ef = 0$ für alle $e \neq f$ aus E , so nennen wir *Primspektrum* $\text{Spec } A$ von A die (bis auf Isomorphie eindeutig bestimmte) Kategorie mit Objektmengen E und Morphismenmengen $\text{Hom}(e, f) = fAe$, deren Komposition von der Multiplikation in A geliefert wird. Ein alter Satz [34] kennzeichnet die idealendlichen Algebren anhand ihrer Primspektren: A ist genau dann *idealendlich*—hat nur endlich viele (zweiseitige) Ideale—wenn jede Endomorphismenalgebra eAe

die Gestalt $k[T]/T^{n(e)}$ hat und jeder Morphismenraum fAe zyklisch ist unter der Aktion von fAf oder von eAe . Letztere Bedingung ist also erfüllt, wenn A darstellungsendlich ist.

1.2. Es sei A idealendlich. Wir nennen zwei Morphismen $\mu, \nu \in fAe$ äquivalent, wenn es Automorphismen ρ und σ gibt derart, dass $\nu = \rho\mu\sigma$. Die Äquivalenzklasse $\vec{\mu}$ von μ nennen wir einen *Strahl* von e nach f . Die Strahlen sind die Morphismen einer Kategorie \vec{A} mit Objektenmenge E , deren Komposition definiert ist vermöge

$$\vec{\mu}\vec{\nu} = \begin{cases} \vec{\mu\nu}, & \text{falls } \{\mu'\nu': \vec{\mu}' = \vec{\mu} \text{ and } \vec{\nu}' = \vec{\nu}\} \text{ ein Strahl ist,} \\ \vec{0}, & \text{sonst.} \end{cases}$$

Es folgt aus der Idealendlichkeit von A , dass \vec{A} eine Strahlenkategorie ist im Sinne der folgenden Definition.

DEFINITION. Eine Kategorie Σ heisst *Strahlenkategorie*, wenn sie Nullmorphismen $y0_x = 0: x \rightarrow y$ besitzt (wobei $0\mu = \mu 0 = 0$ für alle μ gelte), und den folgenden Bedingungen (a)–(e) genügt:

- (a) Verschiedene Objekte von Σ sind nicht isomorph.
- (b) Ist $x \in \Sigma$, so gilt $\text{Hom}(x, y) = \{0\} = \text{Hom}(z, x)$ für fast alle $y, z \in \Sigma$.
- (c) Ist $x \in \Sigma$, so ist $\text{Hom}(x, x)$ für ein geeignetes $n \geq 1$ isomorph zur Halbgruppe $H_n = \{1, \sigma, \sigma^2, \dots, \sigma^{n-1}, \sigma^n = \sigma^{n+1} = \dots = 0\}$.
- (d) Für alle $x, y \in \Sigma$ ist die Menge $\text{Hom}(x, y)$ zyklisch unter der Aktion von $\text{Hom}(x, x)$ oder von $\text{Hom}(y, y)$.
- (e) Für Morphismen λ, μ, ν, π folgt $\nu = \pi$ aus $\mu\nu = \mu\pi \neq 0$ und $\mu = \lambda$ aus $\mu\nu = \lambda\nu \neq 0$.

BEISPIELE. (1) $A = k[T]/T^m$. Dann hat \vec{A} ein Objekt mit Endomorphismenhalbgruppe H_n .

(2) Jede endliche partiell geordnete Menge M kann als Objektenmenge einer Strahlenkategorie aufgefasst werden. Dabei enthält $\text{Hom}(x, y)$ genau dann einen Morphismus $\neq 0$, wenn $x \leq y$. Die Komposition ist die natürliche.

(3) Ist A die Algebra der $n \times n$ -Dreiecksmatrizen, so ist \vec{A} die Strahlenkategorie zur geordneten Menge $\{1 < 2 < \dots < n\}$.

1.3. Einer Strahlenkategorie Σ assoziieren wir die direkte Summe

$$\bigoplus k\Sigma = \bigoplus_{x, y \in \Sigma} k\Sigma(x, y)/k_y 0_x,$$

wobei $k\Sigma(x, y)$ den k -Vektorraum mit Basismenge $\text{Hom}_\Sigma(x, y)$ bezeichnet. Fassen wir die Elemente von $\bigoplus k\Sigma$ als quadratische Matrizen mit Spaltenindex x und Zeilenindex y auf, so liefern die Matrizenmultiplikationsregel und die Komposition von Σ eine Algebrastruktur auf $\bigoplus k\Sigma$. Ist Σ endlich, so hat die Algebra $\bigoplus k\Sigma$ eine Eins und ist idealendlich. Ferner ist Σ dann kanonisch isomorph der Strahlenkategorie zu $\bigoplus k\Sigma$. Insbesondere gibt es zu jedem endlichen Σ ein A mit $\vec{A} \xrightarrow{\sim} \Sigma$.

HAUPTSATZ. Es ist $A \xrightarrow{\sim} \bigoplus k\vec{A}$, falls (a) gilt oder (b):

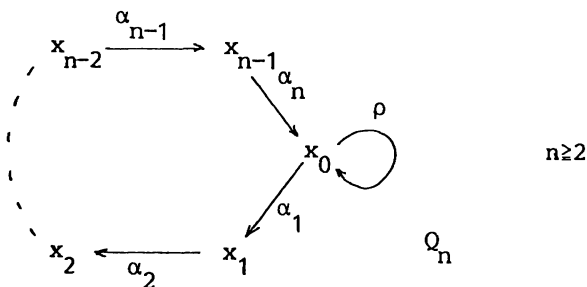
(a) [8] A ist darstellungsendlich und $\text{Char } k \neq 2$.

(b) [8, 15] A ist idealendlich, darstellungsunendlich, und jede echte Restklassenalgebra von A ist darstellungsendlich.

1.4. Die Strahlenkategorien sind die zur Klassifikation der darstellungsendlichen Algebren gesuchten Invarianten. Sie sind kombinatorischer Natur mit leichter algebraischer Schattierung. Wir müssen uns bemühen, sie zu verstehen.

Zunächst bemerken wir, dass jeder Morphismus $\neq 0$ einer Strahlenkategorie Σ eine (im allgemeinen nicht eindeutige) endliche Komposition irreduzibler Morphismen ist (*irreduzibel* bedeute hier nicht null, keine Identität und ohne echte Faktorisierung).

Der Strahlenkategorie Σ ist also ein *Köcher* (= gerichteter Graph) Q_Σ zugeordnet mit den Objekten von Σ als Punkten und den irreduziblen Morphismen als Pfeilen. Umgekehrt gehört zu jedem Köcher Q die *Wegekategorie* WQ : Ihre Objekte sind die Punkte von Q , ihre Morphismen die *Wege* (formale Zusammensetzungen von Pfeilen, darunter die Identitäten), zu denen man formale Nullmorphismen hinzugesellt. Im Fall $Q = Q_\Sigma$ haben wir den kanonischen Funktor $WQ_\Sigma \rightarrow \Sigma$, der Nullmorphismen auf Nullmorphismen abbildet und formale Kompositionen von Pfeilen auf die entsprechenden Kompositionen irreduzibler Morphismen in Σ . Dank diesem Funktor können wir jede Strahlenkategorie als Restklassenkategorie einer Wegekategorie sehen.



FIGUR 1

Ein prominentes Vorbild ist uns der Köcher Q_n von Figur 1. Zu jeder nichtabsteigenden Funktion $f: \{1, \dots, n-1\} \rightarrow \{1, \dots, n\}$ gehört eine grösste Restklassenkategorie R_n^f von WQ , in der die Relationen $\alpha_1 \alpha_n = 0$, $\alpha_n \cdots \alpha_2 \alpha_1 = \rho^2$ und $\alpha_{f i} \cdots \alpha_1 \rho \alpha_n \cdots \alpha_{i+1} = 0$ für $1 \leq i \leq n-1$ erfüllt sind. R_n^f ist eine Strahlenkategorie.

1.5. Allgemein heisse ein irreduzibler Morphismus $\rho: x \rightarrow x$ von Σ (eine Schlaufe von Q_Σ) *kritisch*, wenn ρ^3 nicht null ist und eine Zerlegung $\rho^2 = \alpha_n \cdots \alpha_1$ von ρ^2 in irreduzible Morphismen α_i mit $\alpha_1 \neq \rho$ und $\alpha_1 \rho \alpha_n \neq 0$ existiert.

SATZ [8]. Es sei A darstellungsendlich und S die Menge der kritischen Schleifen in Q_A^- .

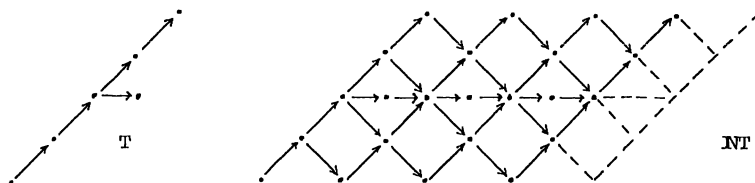
(a) Für jedes $\rho \in S$ ist die Zerlegung $\rho^2 = \alpha_n \cdots \alpha_1$ eindeutig, und die volle Unterkategorie von \vec{A} mit den Spitzen der α_i als Objekten ist isomorph einer Kategorie R_n^f .

(b) Ist ferner $\text{Char } k = 2$, so ist die Anzahl der Isomorphieklassen idealendlicher Basisalgebren B mit $\vec{B} \xrightarrow{\sim} \vec{A}$ gleich der Anzahl Bahnen in 2^S unter der Automorphismengruppe von \vec{A} .

Zum Beispiel gibt es im Fall $f(n-1) \geq 2$ und $\text{Char } k = 2$ zwei nichtisomorphe B mit $\vec{B} \xrightarrow{\sim} R_n^f$, nämlich $B_1 = \bigoplus kR_n^f$ und die Algebra B_2 definiert durch den Köcher Q_n und die Relationen $\alpha_1 \alpha_n = \alpha_1 \rho \alpha_n$, $\alpha_n \cdots \alpha_1 = \rho^2$, $\alpha_{f_i} \cdots \alpha_1 \rho \alpha_n \cdots \alpha_{i+1} = 0$ ($1 \leq i \leq n-1$). Im Fall $\text{Char } k \neq 2$ liefert die Zuordnung $\rho \mapsto \rho(1 - \frac{1}{2}\rho)$, $\alpha_1 \mapsto \alpha_1(1 - \rho)$ und $\alpha_i \mapsto \alpha_i$ für $i \neq 1$ einen Isomorphismus $\bigoplus kR_n^f \xrightarrow{\sim} B_2$.

2. Einfach zusammenhängende Algebren. Nach §1.3 hat jede darstellungsendliche Basisalgebra die Form $\bigoplus k\Sigma$ falls $\text{Char } k \neq 2$. Aber nicht jedes $\bigoplus k\Sigma$ ist darstellungsendlich. Die in §3.4 gegebene Charakterisierung der endlichen Strahlenkategorien Σ mit darstellungsendlichem $\bigoplus k\Sigma$ stützt sich auf gewisse Klassen von Algebren, die wir zuerst vorführen.

2.1. Wir gehen aus von einem endlichen Baum(köcher) T und ordnen ihm einen unendlichen Köcher NT wie folgt zu (Figur 2): Jedem Punkt $t \in T$ entspricht eine Folge von Punkten $(n, t) \in NT$ mit $n \in \mathbb{N}$, jedem Pfeil $s \xrightarrow{\alpha} t$ zwei Folgen $(n, s) \xrightarrow{(n, \alpha)} (n, t)$ und $(n, t) \xrightarrow{\rho(n, \alpha)} (n+1, s)$. Es wirkt auf NT die Verschiebung $\pi: (n, t) \rightarrow (n+1, t)$.



FIGUR 2

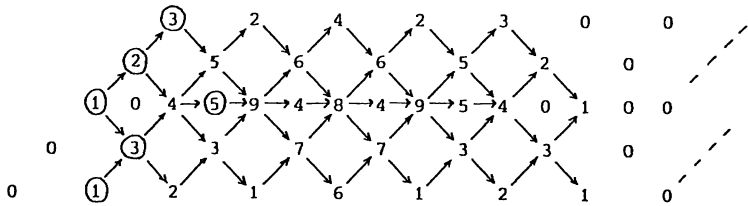
Sei P ein *Muster* von NT , das heie eine Menge von Punkten, die mindestens einen Punkt der Gestalt $(0, t)$ und fr jedes $s \in T$ genau ein Paar der Form (n, s) enthlt. Wir definieren eine \mathbb{N} -wertige *Dimensionsfunktion* $d_P = d$ auf den Punkten von NT so, dass gilt:

(a) $d(x) = 1 + \sum_{y \rightarrow x} d(y)$, falls $x \in P$. Dabei durchluft y die Nocken der Pfeile mit Spitze x .

(b) $d(x) = -d(z) + \sum_{y \rightarrow x} d(y)$, falls $x = \pi z$ und $\sum_{y \rightarrow x} d(y) > d(z) > 0$.

(c) $d(x) = 0$ sonst.

2.2. Sei R_P der volle Unterkcher von NT bestehend aus den Punkten x mit $d(x) > 0$ (Figur 3). Sei $kR_P(x, y)$ der k -Vektorraum, der frei erzeugt wird von den Wegen von x nach y in R_P . Element von $kR_P(x, \pi x)$ ist zum Beispiel die



FIGUR 3

Maschensumme $\sigma_x = \Sigma \rho(\gamma)\gamma$, wobei γ alle Pfeile von R_P mit Nock x durchläuft und $\rho(\rho(n, \alpha)) = (n + 1, \alpha)$ gilt (2.1). Wir bezeichnen mit $k(R_P)(x, y)$ den Restklassenraum von $kR_P(x, y)$ modulo Vektoren der Gestalt $w\sigma_z v$ mit $z \in R_P$, $v \in kR_P(x, z)$ und $w \in kR_P(\pi z, y)$. Die Vektorräume $k(R_P)(x, y)$ sind die Morphismenräume der *Maschenkategorie* $k(R_P)$ von Riedtmann [40], deren Objekte die Punkte von R_P sind, deren Komposition durch die formale Zusammensetzung der Wege induziert wird.

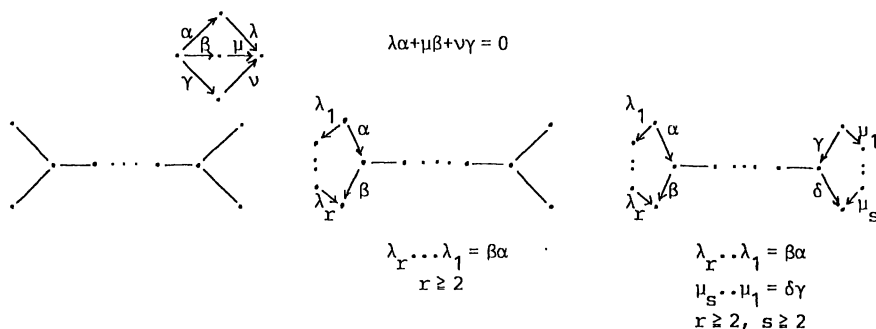
Nun liefert die Multiplikationsregel der Matrizen eine Basisalgebra $A_P = \bigoplus_{p, q \in P} k(R_P)(q, p)$, wobei p der Zeilen- und q der Spalten-index ist. Ferner bestimmt jedes $x \in R_P$ einen A_P -Modul $M(x) = \bigoplus_{p \in P} k(R_P)(p, x)$, dessen Elemente wir als Zeilen auffassen. Die folgenden Aussagen beruhen auf den Eigenschaften der fast zerfallenden Sequenzen von Auslander und Reiten [3, 5, 16, 7, 19]:

- (a) Für jedes $x \in R_P$ ist $M(x)$ unzerlegbar mit Dimension $d_P(x)$.
- (b) $M(x)$ ist projektiv (bzw. injektiv) genau dann, wenn $x \in P$ (bzw. wenn $\pi x \notin R_P$).
- (c) Der Funktor $x \mapsto M(x)$ ist voll treu.
- (d) Wenn R_P endlich ist, und nur dann, erhalten wir mit $x \mapsto M(x)$ eine Bijektion zwischen den Punkten von R_P und den Isomorphieklassen unzerlegbarer A_P -Moduln.

2.3. Für jedes $p \in P$ sei $e(p) \in A_P$ das Element mit $e(p)_{pp} = 1_p \in k(R_P)(p, p)$ als einziger Komponente $\neq 0$. Die $e(p)$ ergeben eine Zerlegung der Eins wie in §1.1 und erlauben uns die Identifizierung von P mit der Objektenmenge von $\text{Spec } A_P$. Ferner ordnen sie jedem endlichdimensionalen A_P -Modul N den *Dimensionsvektor* $\dim N \in \mathbb{Z}^P$ mit $(\dim N)(p) = [Ne(p) : k]$ zu. Ist zum Beispiel $N = k_p$ einfach mit Träger in p , so ist $(\dim N)(q)$ gleich 1 oder 0 je nachdem ob $q = p$ oder $q \neq p$. Wir verstehen \mathbb{Z}^P mit der quadratischen Form

$$\chi_P(\xi) = \sum_{i=0}^{i=2} (-1)^i \sum_{p, q \in P} \xi(p)\xi(q)[\text{Ext}_{A_P}^i(k_p, k_q) : k].$$

SATZ [42, 39, 12]. A_P ist genau dann darstellungsendlich, wenn $\chi_P(\xi) > 0$ für alle $\xi \in \mathbb{N}^P \setminus \{0\}$. In diesem Fall liefert die Abbildung $x \mapsto \dim M(x)$ eine Bijektion zwischen den Punkten von R_P (siehe §2.2) und den $\xi \in \mathbb{N}^P$ mit $\chi_P(\xi) = 1$; ferner gilt dann $[M(x)e(p) : k] \leq 6$ für alle $p \in P$.



FIGUR 4

2.4. Ein A -Modul M heie *omniprsent*, wenn $Me \neq 0$ fr jedes $e \in E$ (§1.1). Und A heie *aufrechtig*, wenn ein omniprsenter unzerlegbarer endlichdimensionaler Modul existiert. Bongartz hat die darstellungsendlichen aufrichtigen Algebren A_P klassifiziert [11, 44]. Er teilt sie ein in 24 unendliche Reihen und 16,815 Ausnahmealgebren kleiner Dimension. Die 24 Reihen spielen eine immense Rolle in den Beweisen der hier zusammengefassten Theorie.

2.5. Happel und Vossieck haben die darstellungsunendlichen Algebren A_P klassifiziert derart, dass $A_P/e(p)$ fr jedes $p \in P$ darstellungsendlich sei. Parallel zu ihnen ist Bongartz auf dieselbe Algebrenliste gestossen, aber mit einer anderen Charakterisierung, die in §3.4 bentigt wird. Die Liste umfasst die Algebren von Figur 4, die wir durch Kcher und Relationen beschreiben (Kanten drfen beliebig orientiert werden). Darberhinaus enthlt sie 4,299 Ausnahmealgebren, von Happel und Vossieck sehr bersichtlich in 141 Familien eingeordnet [13, 31, 14, 8, 44]. Von Hhne konnte die Liste 1985 ohne Computerhilfe besttigen [49].

3. Das Endlichkeitskriterium von Bongartz. Es bezeichne Σ stets eine Strahlenkategorie.

3.1. Eine *Darstellung* von Σ ber k ist ein kontravarianter Funktor V von Σ in die Kategorie der k -Vektorrume derart, dass $V({}_y0_x)$ null sei fr alle $x, y \in \Sigma$. Die Darstellungen von Σ bilden eine abelsche Kategorie, die im Falle eines endlichen Σ vermge $V \mapsto \bigoplus_{y \in \Sigma} V(y)$ quivalent ist zur Kategorie der $\bigoplus k\Sigma$ -Moduln. Insbesondere sind die Begriffe "direkte Summe" und "unzerlegbar" erklrt. Die *Dimension* einer Darstellung V ist $\Sigma_y[V(y) : k]$, und Σ heit (lokal) *darstellungsendlich*, wenn es fr jedes $y \in \Sigma$ nur endlich viele Isomorphieklassen unzerlegbarer V mit $V(y) \neq 0$ gibt.

SATZ [8, 41]. Eine idealendliche Basisalgebra A ist genau dann darstellungsendlich, wenn ihre Strahlenkategorie \vec{A} es ist.

3.2. Alle Algebren der Liste von Bongartz-Happel-Vossieck (§2.5) mit Ausnahme der ersten sind idealendlich. Die dazu assoziierten und alle isomorphen Strahlenkategorien nennen wir *kritisch*.

Ein *kritisches Diagramm* von Σ ist ein Funktor $D: K \rightarrow \Sigma$ mit den Eigenschaften (a) bis (c): (a) K ist kritisch. (b) Es ist $D\mu = 0$ genau dann, wenn $\mu = 0$. (c) Ist in $x \xrightarrow{\mu} y \xleftarrow{\alpha} z$ oder $x \xleftarrow{\mu} y \xrightarrow{\alpha} z$ der Morphismus α von K irreduzibel und faktorisiert $D\mu$ durch $D\alpha$, so faktorisiert auch μ durch α .

Wir nennen $D: K \rightarrow \Sigma$ *konvex*, wenn für jeden Morphismus $0 \neq \mu: x \rightarrow y$ von K und jede Zerlegung $D\mu = \beta_n \cdots \beta_1$ von $D\mu$ in irreduzible Morphismen von Σ Morphismen α_i mit $D\alpha_i = \beta_i$ und $\mu = \alpha_n \cdots \alpha_1$ existieren.

Schliesslich nennen wir *Kette* von Σ eine *unendliche* alternierende Folge $\cdot \xrightarrow{\rho_1} \cdot \xleftarrow{\sigma_1} \cdot \xrightarrow{\rho_2} \cdot \xleftarrow{\sigma_2} \cdot \dots$ von Morphismen derart, dass für jedes i keine der Gleichungen

$$\rho_i = \sigma_i \xi, \quad \rho_i \xi = \sigma_i, \quad \xi \sigma_i = \rho_{i+1}, \quad \sigma_i = \xi \rho_{i+1}$$

eine Lösung ξ in Σ hat.

SATZ [13, 14, 8, 22]. *Die folgenden Aussagen sind äquivalent:*

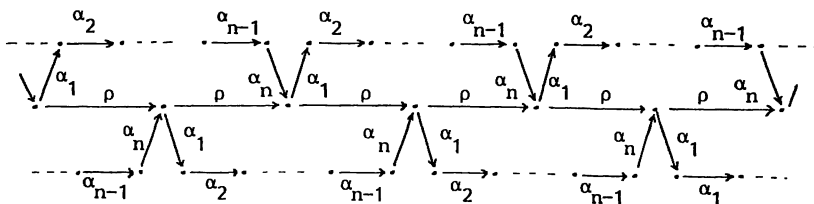
- (i) Σ ist darstellungsendlich.
- (ii) In Σ existiert keine Kette und kein kritisches Diagramm.
- (iii) In Σ existiert keine Kette und kein konvexes kritisches Diagramm.

Die ohne "BHV-Liste" erfolglose Suche nach einem Kriterium dieser Art war ein Leitmotiv der Fünfziger Jahre (Jans, Yoshii).

3.3. So elementar die Formulierung von §3.2 auch sein mag, der Satz ist unehrlich und nutzlos: unehrlich, weil er seine Quellen verdeckt, nutzlos, weil das kombinatorische Dickicht von Σ mit (ii) und (iii) kaum zu durchdringen ist. Wir müssen höher steigen.

DEFINITION. Eine *Ueberlagerung* von Σ ist ein Funktor $F: \Sigma' \rightarrow \Sigma$ mit den Eigenschaften (a) und (b): (a) Σ' ist eine Strahlenkategorie, und es ist $F\mu' = 0$ genau dann, wenn $\mu' = 0$. (b) Zu jedem Morphismus $\mu: x \rightarrow y$ ungleich 0 in Σ und jedem $x' \in \Sigma'$ mit Bild x (bzw. jedem $y' \in \Sigma'$ mit Bild y) gibt es genau ein μ' mit Quelle x' (bzw. Ziel y') und Bild μ .

Unsere Definition löst die üblichen Routine-Ueberlagerungen aus: ein Spezialfall der Ueberlagerungstheorie semisimplizialer Komplexe. Insbesondere gehören zu jedem "zusammenhängenden" Σ eine "*universelle*" Ueberlagerung $F_\Sigma: \tilde{\Sigma} \rightarrow \Sigma$ und eine *Fundamentalgruppe* Π_Σ mit $\tilde{\Sigma}/\Pi_\Sigma \xrightarrow{\sim} \Sigma$. Zum Beispiel hat \tilde{R}_n^f als Köcher \tilde{Q}_n (Figur 5) und wird durch Relationen bestimmt, die gleich lauten wie bei R_n^f (§1.4).



FIGUR 5

SATZ [16, 18, 8, 22]. *Enthält Σ keine Kette, so gilt:*

(a) *Die Fundamentalgruppe Π_Σ ist frei (nicht kommutativ).*

(b) *Von einem Punkt \tilde{x} zu einem Punkt \tilde{y} führen nur endlich viele Wege des Köchers $Q_{\tilde{\Sigma}}$: die Kompositionen dieser Wege in $\tilde{\Sigma}$ sind alle gleich.*

Es folgt insbesondere, dass $Q_{\tilde{\Sigma}}$ keine geschlossenen Wege enthält (ausser den identischen), und dass es zwischen zwei Objekten von $\tilde{\Sigma}$ höchstens einen Morphismus $\neq 0$ gibt.

3.4. KRITERIUM VON BONGARTZ [13, 14, 8, 22]. *Die Strahlenkategorie Σ ist genau dann darstellungsendlich, wenn $\tilde{\Sigma}$ keine Kette und keine konvex kritische Unterkategorie enthält.*

Dabei heisst eine Unterkategorie von Σ *konvex kritisch*, wenn der Einbettungsfunktor konvex, kritisch und voll treu ist (§3.2).

Das Kriterium erweist sich in der Praxis [21] als sehr handlich (insbesondere bei R_n^f), weil "konvexe" Unterkategorien von $\tilde{\Sigma}$ leicht mittels $Q_{\tilde{\Sigma}}$ zu beschreiben und kritische Strahlenkategorien leicht anhand auftretender kommutativer Polygone zu identifizieren sind.

3.5. SATZ [37, 6, 15]. *Ist die Algebra A darstellungsunendlich, so ist die Anzahl der Dimensionen, für die es unendlich viele Isomorphieklassen unzerlegbarer A -Moduln gibt, auch selbst unendlich.*

Der Satz geht auf eine um die 40 Jahre alte Vermutung zurück, die Brauer und Thrall zugeschrieben wird. Es folgt aus ihm, dass jedes Element einer unendlichen irreduziblen algebraischen Familie generisch nicht-isomorpher Algebren-Strukturen auf k^n darstellungsunendlich ist [23, 24]. Einer der möglichen Beweise des Satzes ergibt sich leicht aus der Reduktion der Aussage auf kritische Kategorien vermöge §1.3 (b), §3.4 und [25, 35].

4. Ausblick. Es ist unsere Meinung, dass die darstellungsendlichen Algebren den ausgereiften und technisch schwierigeren Teil der heutigen Darstellungstheorie endlichdimensionaler Algebren ausmachen. Ihre Erforschung ist das Werk vieler. Sucht man dennoch nach unerwarteten, anspruchsvollen und entscheidenden Einzelleistungen, so ragen die Namen Bongartz und Roiter heraus.

Wir müssen aber auch einsehen, dass die zuerst eroberte Stellung auch am weitesten entfernt vom Zentrum der Algebra liegt. Es ist zu erwarten, dass nun die Festung der zahmen Algebren systematischer eingeschlossen wird. Als Beispiel systematischer Arbeit sehen wir die Charakterisierung der zahmen vervollständigten geordneten Mengen durch Nazarova und Roiter [38, 46]. Zitieren müssen wir auch die Arbeit [20] von Dowbor und Skowronski, die frühere Gedankengänge von I. M. Gel'fand und Ponomarev [27] überlagerungstheoretisch interpretiert und weiterführt, wobei die Beweise den Bereich der endlichdimensionalen Darstellungen sprengen. Leider fehlt uns die Einsicht in die weiteren Schubladen der Konkurrenz.

Die Mehrheit der neuesten Publikationen wird motiviert durch das Bedürfnis nach besseren Verbindungen zum Festland der Mathematik. Der Brücke zur Kommutativen Algebra ist der Bericht M. Auslanders gewidmet [4]. Andere Publikationen kreisen um den Begriff der derivierten Kategorien von Grothendieck-Verdier [48]. Konzipiert wurde dieser Begriff zur Formulierung der Dualitätstheorie von Grothendieck-Roos. In anderem Gewand taucht diese Theorie heute als "Kipptheorie" auf, initiiert von I. M. Gel'fand und ausgebaut hauptsächlich von Brenner-Butler [17] und Happel-Ringel [30]. Interessant ist dabei, dass sehr verschiedenartige abelsche Kategorien äquivalente derivierte Kategorien besitzen. So produzieren die kohärenten Garben auf Grassmannschen Varietäten und Quadriken "dieselben" derivierten Kategorien wie geeignete endlichdimensionale Algebren [10, 9, 32, 28, 33]. Happel beschreibt in [29] die Algebren, die ihre derivierten Kategorien mit den Wegealgebren der Dynkinköcher "teilen." In dieselbe Richtung weisen die Arbeiten [1, 47, 29] über triviale Algebrenweiterungen. Schliesslich zeigen Geigle und Lenzing [26], dass die kanonischen Algebren Ringels [44] die gleichen derivierten Kategorien liefern wie gewisse graduierte Modulgarben auf der projektiven Geraden; so verbinden sie Atiyahs Arbeit über Vektorbündel auf elliptischen Kurven [2] mit Ringels Klassifikation der Moduln über zahmen "tubularen" Algebren.

BIBLIOGRAPHIE

1. I. Assem, D. Happel, and O. Roldán, *Representation-finite trivial extension algebras*, J. Pure Appl. Algebra **33** (1984), 235–242.
2. M. F. Atiyah, *Vector bundles over an elliptic curve*, Proc. London Math. Soc. (3) **7** (1957), 414–452.
3. M. Auslander, *Applications of morphisms determined by objects*, Representation Theory of Algebras (R. Gordon, ed.), Lecture Notes in Pure and Appl. Math. **37** (1976), 245–327.
4. —, *The what, where and why of almost split sequences*, Proc. Internat. Congr. Math. (Berkeley, Calif., U.S.A., 1986).
5. M. Auslander and I. Reiten, *Representation theory of artin algebras*. III–V, Comm. Algebra **3** (1975), 239–294; **5** (1977), 443–518; **5** (1977), 519–554.
6. R. Bautista, *On algebras of strongly unbounded representation type*, Comment. Math. Helv. **60** (1985), 392–399.
7. R. Bautista, F. Larrión, and L. Salmerón, *On simply connected algebras*, J. London Math. Soc. **27** (1983), 212–220.
8. R. Bautista, P. Gabriel, A. Roiter, and L. Salmerón, *Representation-finite algebras and multiplicative bases*, Invent. Math. **81** (1985), 217–285.
9. A. A. Beilinson, *Kohärente Garben auf P^n und Probleme der Linearen Algebra*, Funktsional. Anal. i Prilozhen. **12** (1978), 68–69. (Russisch)
10. I. N. Bernstein, I. M. Gelfand, and S. I. Gelfand, *Algebraische Bündel auf P^n und Aufgaben der Linearen Algebra*, Funktsional. Anal. i Prilozhen. **12** (1978), 66–67. (Russisch)
11. K. Bongartz, *Treue einfach zusammenhängende Algebren*. I, Comment. Math. Helv. **57** (1982), 282–330.
12. —, *Algebras and quadratic forms*, J. London Math. Soc. **28** (1983), 461–469.
13. —, *A criterion for finite representation type*, Math. Ann. **269** (1984), 1–12.
14. —, *Critical simply connected algebras*, Manuscripta Math. **46** (1984), 117–136.
15. —, *Indecomposable modules are standard*, Comment. Math. Helv. **60** (1985), 400–410.

16. K. Bongartz and P. Gabriel, *Covering spaces in representation theory*, Invent. Math. **65** (1982), 331–378.
17. S. Brenner and M. C. R. Butler, *Generalizations of the Bernstein-Gelfand-Ponomarev reflection functors*, Representation Theory. II, Lecture Notes in Math., vol. 832, Springer, 1980, pp. 103–169.
18. O. Betscher and P. Gabriel, *The standard form of a representation-finite algebra*, Bull. Soc. Math. France **111** (1983), 21–40.
19. B. Conti, *Simply connected algebras of tree-class A_n and D_n* , Representation Theory. I, Lecture Notes in Math., vol. 1177, Springer, 1986, pp. 60–90.
20. P. Dowbor and A. Skowronski, *Galois coverings of representation-finite algebras*, Comment. Math. Helv. (erscheint demnächst).
21. U. Fischbacher, *The representation-finite algebras with three simple modules*, Representation Theory. I, Lecture Notes in Math., vol. 1177, Springer, 1986, pp. 94–114.
22. —, *Zur Kombinatorik der Algebren mit endlich vielen Idealen*, J. Reine Angew. Math. (erscheint demnächst).
23. P. Gabriel, *Finite representation type is open*, Representation Theory of Algebras, Lecture Notes in Math., vol. 488, Springer, 1975, pp. 132–155.
24. —, *Auslander-Reiten sequences and representation-finite algebras*, Representation Theory. I, Lecture Notes in Math., vol. 831, Springer, 1980, pp. 1–71.
25. —, *The universal cover of a representation-finite algebra*, Representations of Algebras, Lecture Notes in Math., vol. 903, Springer, 1981, pp. 68–105.
26. W. Geigle and H. Lenzing, *A class of weighted projective curves arising in representation theory of finite-dimensional algebras*, Preprint, 1986.
27. I. M. Gelfand and V. A. Ponomarev, *Unzerlegbare Darstellungen der Lorentzgruppe*, Uspekhi Mat. Nauk **23** (1968), 3–60. (Russisch)
28. S. I. Gelfand, *Garben auf P_n und Aufgaben der linearen Algebra*, Anhang zur russischen Fassung von "Vector bundles on complex projective spaces" von Okonek, Schneider und Spindler, "Mir", Moskau, 1984, pp. 278–305.
29. D. Happel, *On the derived category of a finite-dimensional algebra*, Comment. Math. Helv. (erscheint demnächst).
30. D. Happel and C. M. Ringel, *Tilted algebras*, Trans. Amer. Math. Soc. **274** (1982), 399–443.
31. D. Happel and D. Vossieck, *Minimal algebras of infinite representation type with preprojective component*, Manuscripta Math. **42** (1983), 221–243.
32. M. M. Kapranov, *The derived categories of coherent sheaves on grassmannians*, Functional Anal. Appl. **17** (1983), 78–79. (Russisch)
33. —, *The derived category of coherent sheaves on a quadric*, Functional Anal. Appl. **20** (1986), 67. (Russisch)
34. H. Kupisch, *Symmetrische Algebren mit endlich vielen unzerlegbaren Darstellungen*, I, J. Reine Angew. Math. **249** (1965), 1–25.
35. R. Martinez and J. A. de la Peña, *Automorphisms of representation-finite algebras*, Invent. Math. **72** (1983), 359–362.
36. K. Morita, *Duality for modules and its application to the theory of rings with minimum condition*, Sci. Rep. Tokyo Kyoiku Daigaku **6** (1958/59), 83–142.
37. L. A. Nazarova and A. V. Roiter, *Categorical matrix problems and the Brauer-Thrall conjecture* (deutsche Fassung: Mitt. Math. Sem. Giessen **115** (1975), 1–153).
38. —, *Darstellungen und Formen schwach vervollständigter partiell geordneter Mengen*, Lineare Algebra und Darstellungstheorie, Kiev, 1983, pp. 19–54. (Russisch)
39. S. A. Ovsienko, *Ganzzahlige schwach positive Formen*, Schursche Matrizenprobleme und quadratische Formen, Kiev, 1978, pp. 3–17. (Russisch)
40. Chr. Riedtmann, *Algebren, Darstellungsköcher, Ueberlagerungen und zurück*, Comment. Math. Helv. **55** (1980), 199–224.
41. —, *Algèbres de type de représentation fini*, Séminaire Bourbaki **37** (1985), 335–350.
42. C. M. Ringel, *Representations of K -species and bimodules*, J. Algebra **41** (1976), 269–302.

43. —, *Indecomposable representations of finite-dimensional algebras*, Proc. Internat. Congr. Math. (Warszawa, 1983), vol. 1, 425–436.
44. —, *Tame algebras and quadratic forms*, Lecture Notes in Math., vol. 1099, Springer, 1984.
45. A. V. Roiter, *Verallgemeinerung eines Satzes von Bongartz*, Akad. Nauk Ukrain. SSR Inst. Mat. Preprint 1981, no. 17, pp. 1–32. (Russisch)
46. —, *Representations of posets and tame matrix problems*, Proc. Durham Conf. on Representations of Algebras and Groups (erscheint demnächst).
47. H. Tachikawa and T. Wakamatsu, *Tilting functors and stable equivalences for self-injective algebras*, Preprint.
48. J.-L. Verdier, *Catégories dérivées, état 0*, Séminaire de Géométrie Algébrique 4 1/2, Lecture Notes in Math., vol. 569, Springer, 1977, pp. 262–311.
49. H.-J. von Hoehne, *Ganze quadratische Formen und Algebren*, Dissertation, Freie Universität Berlin, August 1986, pp. 1–122.

ABSTRACT. *Representations of finite-dimensional algebras.* Let A be a finite-dimensional algebra (associative, with 1) over an algebraically closed field k . The objective is to classify the finite-dimensional A -modules. Efficient algorithms are available if A has only finitely many indecomposable modules up to isomorphism. This report is mainly devoted to the reduction of the classification of such algebras to a combinatorial problem. A complete English version of the German original is available at the author's address.

РЕЗЮМЕ. *Представления конечномерных алгебр.* Пусть A конечномерная алгебра (ассоциативная, с 1) над алгебраически замкнутым полем k . Цель — классификация конечномерных A -модулей. Эффективные алгоритмы находятся в нашем распоряжении, если A имеет только конечное число неразложимых модулей с точностью до изоморфности. Этот доклад главным образом посвящен сведению классификации таких алгебр комбинаторной проблеме. Полное русское изложение сей немецкой статьи можно получить у автора.

MATHEMATISCHES INSTITUT, RÄMISTRASSE 74, CH-8001 ZÜRICH, SWITZERLAND

Milnor K -Theory and Galois Cohomology

A. S. MERKURJEV

Let F be a field and m a natural number, $(m, \text{char } F) = 1$. It follows from the exact sequence $1 \rightarrow \mu_m \rightarrow F_{\text{sep}}^* \xrightarrow{m} F_{\text{sep}}^* \rightarrow 1$ and the Hilbert Theorem 90 that there is the natural isomorphism $l: F^*/F^{*m} \rightarrow H^1(F, \mu_m)$. The cup-product in the cohomology theory of groups gives the homomorphism

$$F^* \otimes F^* \otimes \cdots \otimes F^* \rightarrow H^n(F, \mu_m^{\otimes n}),$$

satisfying the Steinberg condition and therefore giving the homomorphism:

$$\alpha_{n,m}: K_n(F)/mK_n(F) \rightarrow H^n(F, \mu_m^{\otimes n}),$$

by the formula [10]: $\alpha_{n,m}(\{a_1, a_2, \dots, a_n\}) = l(a_1) \cup l(a_2) \cup \cdots \cup l(a_n)$, where $K_n(F)$ is the Milnor group of F . This homomorphism is called the norm residue map of degree n .

There is a conjecture that all $\alpha_{n,m}$ are isomorphisms for any F . If $n = 1$ it is a consequence of the Hilbert Theorem 90. In the case $n = 2$ this conjecture was proved in [5]. If $n \geq 3$ the answer is known only for some fields.

The theorem stating the bijectivity of $\alpha_{2,m}$ has many applications. The present paper is devoted to some applications in the Brauer group theory.

If $\mu_m \subset F$, then

$$H^2(F, \mu_m^{\otimes 2}) = H^2(F, \mu_m) \otimes \mu_m = {}_m\text{Br}(F) \otimes \mu_m,$$

where ${}_m\text{Br}(F)$ is the subgroup of elements of exponent m in the Brauer group $\text{Br}(F)$ of F . In this case the norm residue map of degree 2 coincides with the homomorphism

$$K_2(F)/mK_2(F) \rightarrow {}_m\text{Br}(F) \otimes \mu_m,$$

introduced in [9].

1. Cyclic algebras. Let F be a field and G the Galois group of F_{sep}/F . We denote by $X(F)$ the character group $\text{Hom}_c(G, \mathbf{Q}/\mathbf{Z})$ of G . For any character $\chi \in X(F)$ let $F(\chi)/F$ be the field extension corresponding to $\ker \chi \in G$ by the Galois theory. It is clear that $F(\chi)/F$ is the cyclic extension of degree n which equals the order of χ in $X(F)$. The Galois group of this extension has the natural generator σ_χ , uniquely determined by the equality $\chi(\sigma_\chi) = 1/n + \mathbf{Z}$.

Let $a \in F^*$, $\chi \in X(F)$, $n = (F(\chi): F)$. We define the structure of an F -algebra on an n -dimensional vector space $A = A(a, \chi)$ over $F(\chi)$ with basis u_0, u_1, \dots, u_{n-1} by the following formula:

$$(xu_i) \cdot (yu_j) = \begin{cases} x \cdot \sigma^i(y) \cdot u_{i+j}, & i+j < n, \\ a \cdot x \cdot \sigma^i(y) \cdot u_{i+j-n}, & i+j \geq n. \end{cases}$$

$A(a, \chi)$ is a central simple algebra over F of dimension n^2 and is said to be a cyclic algebra of degree n [1].

There is a conjecture that $\text{Br}(F)$ is generated by the classes $[A(a, \chi)]$ of cyclic algebras. We can formulate this conjecture more precisely: for any $m \in \mathbb{N}$ the group ${}_m\text{Br}(F)$ is generated by $A(a, \chi)$, $a \in F^*$, $\chi \in {}_mX(F)$.

In the following cases the last conjecture has been proved:

1. If $(m, \text{char } F) = 1$ and $\mu_m \subset F$. In this case the conjecture is equivalent to the surjectivity of the norm residue map $\alpha_{2,m}$. Indeed, the element $\alpha_{2,m}(\{a, b\})$ equals $[A(a, \chi_b)] \otimes \xi_m \in {}_m\text{Br}(F) \otimes \mu_m$, where χ_b is the unique character, such that $F(\chi_b) = F(b^{1/m})$, $\sigma_\chi(b^{1/m}) = \xi_m \cdot b^{1/m}$, and any character $\chi \in {}_mX(F)$ equals χ_b for some $b \in F^*$.

2. If m is a prime number and $(F(\mu_m): F) \leq 3$. In particular, the conjecture is proved if $m = 2, 3$ [5, 6].

3. If $m = 4$ [5].

4. If $m = p^k$ and $p = \text{char } F$ [1].

There are the following relations in $\text{Br}(F)$ between the classes of cyclic algebras:

- (1) $[A(ab, \chi)] = [A(a, \chi)] + [A(b, \chi)]$,
- (2) $[A(a, \chi + \rho)] = [A(a, \chi)] + [A(a, \rho)]$,
- (3) $[A(a, \chi)] = 0$ if a is a norm in $F(\chi)/F$.

There is a conjecture that any relation between the classes $[A(a, \chi)]$ with $a \in F^*$, $\chi \in {}_mX(F)$ in ${}_m\text{Br}(F)$ is a consequence of (1)–(3). This conjecture has been proved in the following cases:

1. $(m, \text{char } F) = 1$ and $\mu_m \subset F$. In this case the conjecture is equivalent to the injectivity of the norm residue map $\alpha_{2,m}$.

2. If m is a prime number and $(F(\mu_m): F) \leq 2$. In particular, the conjecture is proved if $m = 2, 3$.

3. If $m = p^k$ and $p = \text{char } F$.

We can take both conjectures together and consider the following version of the norm residue map $\alpha_{2,m}$. Let $S_m(F)$ be the factorgroup $(F^* \otimes {}_mX(F))/B$ where B is the subgroup generated by $a \otimes \chi$, where a is a norm in $F(\chi)/F$.

The correspondence $a \otimes \chi \mapsto [A(a, \chi)]$ defines the following homomorphism: $\beta_m: S_m(F) \rightarrow {}_m\text{Br}(F)$. The first conjecture is equivalent to the surjectivity of β_m and the second conjecture to the injectivity of β_m .

If $\mu_m \subset F$, then the homomorphism $\beta_m \otimes \mu_m$ is equal to the norm residue map $\alpha_{2,m}$ and therefore is an isomorphism.

One can show that to prove the bijectivity of β_m it is sufficient to construct the corestriction map $S_m(L) \rightarrow S_m(F)$ for a finite extension L/F . So far we can

define such a map only for some quadratic extensions. This gives the proof in the case $(F(\mu_m): F) \leq 2$.

2. Generators and relations in Brauer group. The problem is to find a set of generators and relations in $\text{Br}(F)$. It is sufficient to study the p -primary component $\text{Br}(F)\{p\}$ of $\text{Br}(F)$ for any prime p . We assume that $p \neq \text{char } F$.

Let $k \in \mathbb{N}$, $q = p^k$, $F_k = F(\mu_q)$. For $x \in F_k^*$ and $\chi \in {}_qX(F_k)$ let $[x, \chi] = \text{cor}([A(x, \chi)]) \in {}_q\text{Br}(F)$, where $\text{cor}: \text{Br}(F_k) \rightarrow \text{Br}(F)$ is a corestriction map [3]. We know that ${}_q\text{Br}(F_k)$ is generated by $[A(x, \chi)]$. There is a question whether ${}_q\text{Br}(F)$ is generated by $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F_k)$.

A simple example: $F = \mathbb{R}$, $p = 2$, $k \geq 2$ shows that the answer is negative in general. Unexpectedly the answer is positive in almost all other cases.

We suppose that F_k/F is a cyclic extension (if, for example, p is odd).

The following case we call exceptional: $p = 2$, F_k/F is a quadratic extension, and $\tau(\xi_q) = \xi_q^{-1}$, where τ is a generator of $\text{Gal}(F_k/F)$. (The example above belongs to the exceptional case.) The case which is not exceptional is called general. When p is odd or $\sqrt{-1} \in F$, $p = 2$ we have the general case.

THEOREM [7]. 1. *In the general case ${}_q\text{Br}(F)$ is generated by $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F_k)$.*

2. *In the exceptional case ${}_q\text{Br}(F)$ is generated by $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F_k)$ with the classes of quaternion algebras*

$$\langle a, \rho \rangle \stackrel{\text{def}}{=} [A(a, \rho)], \quad a \in F^*, \rho \in {}_2X(F).$$

If F_k/F is not a cyclic extension (this is possible only if $p = 2$) the situation is more complicated. There are some arguments in [7], showing that in this case the group ${}_q\text{Br}(F)$ does not have a "simple set" of generators.

We have the following relations between $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F_k)$:

- (1) $[xy, \chi] = [x, \chi] + [y, \chi]$,
- (2) $[x, \chi + \eta] = [x, \chi] + [x, \eta]$,
- (3) $[x, \chi] = 0$ if x is a norm in $F_k(\chi)/F_k$,
- (4) $[\tau(x), \tau(\chi)] = [x, \chi]$, where $\tau \in \text{Gal}(F_k/F)$.

THEOREM [7]. *In the general case any relation between $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F)$ is a consequence of (1)–(4).*

In the exceptional case we have the following additional relations between $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F_k)$, and the classes of quaternion algebras $\langle a, \eta \rangle$, $a \in F^*$, $\eta \in {}_2X(F)$:

- (5) $\langle ab, \eta \rangle = \langle a, \eta \rangle + \langle b, \eta \rangle$,
- (6) $\langle a, \eta + \rho \rangle = \langle a, \eta \rangle + \langle a, \rho \rangle$,
- (7) $\langle a, \rho \rangle = 0$ if a is a norm in $F(\rho)/F$.
- (8) $[x, \chi_{F_k}] = \langle N_{F_k/F}(x), \chi \rangle$, $x \in F_k^*$, $\chi \in {}_2X(F)$.
- (9) $[a, \chi] = 0$, where $a \in F^*$ and $F_k(\chi) = F_k(b^{1/m})$ for some $b \in F^*$.

THEOREM [7]. *In the exceptional case any relation between $[x, \chi]$, $x \in F_k^*$, $\chi \in {}_qX(F_k)$, and $\langle a, \eta \rangle$, $a \in F^*$, $\eta \in {}_2X(F)$, is a consequence of (1)–(9).*

Studying our generators $[x, \chi]$ in detail we can prove the following statement:

THEOREM [7]. 1. *In the general case ${}_q\text{Br}(F)$ is generated by the classes of algebras of dimension not more than $q^{2p^{k-1}}$.*

2. *In the exceptional case ${}_q\text{Br}(F)$ is generated by the classes of algebras of dimension not more than q .*

COROLLARY [6]. *The group ${}_p\text{Br}(F)$ is always generated by the classes of algebras of dimension p^2 .*

3. Brauer group as an abstract group. It is a well-known fact that any Brauer group is abelian and torsion. In [2] A. Brumer and M. Rosen conjectured that if $2 \cdot \text{Br}(F)\{p\} \neq 0$ then $\text{Br}(F)\{p\}$ contains a nontrivial divisible subgroup.

THEOREM [6]. *Let p be an odd number, $(F(\xi_p): F) \leq 3$, and $\text{Br}(F)\{p\} \neq 0$. Then $\text{Br}(F)\{p\}$ contains a nontrivial divisible subgroup.*

This theorem proves the conjecture of A. Brumer and M. Rosen for $p = 3$ and any field.

In the case $p = 2$ a nontrivial group $\text{Br}(F)\{p\}$ may not contain a divisible subgroup. For example, any elementary abelian 2-group is isomorphic to $\text{Br}(F)\{2\}$ for some field F . The situation is different if $\text{Br}(F)$ contains an element of order 4.

THEOREM [6]. *If $2 \cdot \text{Br}(F)\{2\} \neq 0$, then $\text{Br}(F)\{2\}$ contains a nontrivial divisible subgroup.*

In [4] B. Fein and M. Schacher conjectured that any divisible abelian torsion group is a Brauer group for some field. This conjecture is proved in [8].

REFERENCES

1. A. A. Albert, *Structure of algebras*, Amer. Math. Soc. Colloq. Publ., vol. 24, Amer. Math. Soc., Providence, R.I. 1961.
2. A. Brumer and M. Rosen, *On the size of the Brauer group*, Proc. Amer. Math. Soc. **19** (1968), 707–711.
3. P. K. Draxl, *Skew fields*, London Math. Soc. Lecture Note Ser., vol. 81, Cambridge Univ. Press, 1982.
4. B. Fein and M. Schacher, *Brauer groups of fields*, Ring Theory and Algebra. III, Lecture Notes in Pure and Appl. Math., vol. 55, Dekker, New York, 1980, pp. 345–356.
5. A. S. Merkurjev and A. A. Suslin, *K-cohomology of Severi-Brauer varieties and the norm residue homomorphism*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), 307–340 = Math. USSR Izv. **21** (1983).
6. A. S. Merkurjev, *Brauer groups of fields*, Comm. Algebra **11** (1983), 2611–2624.

7. —, *On the structure of the Brauer group of fields*, Izv. Akad. Nauk SSSR Ser. Mat. **49** (1985), 828–846=Math. USSR Izv. **27** (1986), 141–157.
8. —, *Divisible Brauer groups*, Uspekhi Mat. Nauk **40** (1985), no. 2 (242), 213–214.
9. J. Milnor, *Introduction to algebraic K-theory*, Princeton Univ. Press, Princeton, N.J., 1971.
10. J. Tate, *Relations between K_2 and Galois cohomology*, Invent. Math. **36** (1976), 257–274.

LENINGRAD UNIVERSITY, LENINGRAD 119164, USSR

Modern Developments in Invariant Theory

V. L. POPOV

Invariant theory, which is almost one and a half centuries old, is now going through a new period of vigorous development: a satisfactory comprehension of a general picture within the classical situation has been attained and the new important directions, methods, and applications are being discovered. Now, as a matter of fact, invariant theory is identified with the theory of algebraic group actions on algebraic varieties (schemes) whose local aspect (concerned with the actions on affine varieties) is, on the one hand, the most immediately connected with the classical understanding of invariant theory and, on the other, the most elaborated.

I have to restrict myself here mostly to some results of principal importance that form a group around the classical theme of description of invariants and covariants, and I hope that the addresses by C. de Concini and W. Borho will give an idea of some other up-to-date aspects and applications of invariant theory.

1. Classical problems. Let k be an algebraically closed field, V a finite-dimensional vector space over k , $k[V]$ the algebra of regular functions (polynomials) on V , and G an algebraic subgroup of $\mathrm{GL}(V)$. G acts on $k[V]$.

The main problem of the classical theory is to give "the explicit description" of the algebra of invariants $k[V]^G$. This presupposes (a) solving the question of whether $k[V]^G$ is finitely generated over k and, if yes, (b) giving a constructive way to find explicitly a minimal system of homogeneous generators of $k[V]^G$ (or, what is ideally desirable, even to exhibit such a system actually).

The problem of describing all G such that $k[V]^G$ is finitely generated is known as the original Hilbert's fourteenth problem. (In fact Hilbert's formulation at the 1900 Paris Congress was somewhat different.) In his address at the 1958 Edinburgh Congress, M. Nagata gave an example of such G that $k[V]^G$ is not finitely generated. Actually it is more natural to consider a general situation when G rationally acts on an arbitrary affine (i.e., finitely generated and reduced) k -algebra A or, geometrically, when G algebraically acts on an affine algebraic variety X with the coordinate algebra $k[X] = A$, see [1, 2]. M. Nagata [1]

considered the problem in this general setting and called it "the generalized fourteenth problem." It is completely solved now:

THEOREM 1. *The following properties of G are equivalent: (a) for all affine k -algebras A on which G acts rationally by k -algebra automorphisms A^G is also affine; (b) G is reductive (i.e., its unipotent radical is trivial).*

The proof (b) \Rightarrow (a) goes back to Hilbert when $\text{char } k = 0$; it is much more subtle when $\text{char } k > 0$ and follows from M. Nagata [28] and W. Haboush [29] (see also [2]). The proof (a) \Rightarrow (b) was given by the author [3].

The original Hilbert's fourteenth problem, as distinct from the generalized one, is not yet solved. Its important special feature is that one can extend the action $G: V$ up to an action of a reductive group which contains G (namely, $\text{GL}(V)$). One can use this as follows. Let R be an arbitrary linear algebraic group, H its closed subgroup, and A a k -algebra on which R rationally acts by k -algebra automorphisms. Then k -algebras $(A \otimes k[R/H])^R$ and A^H are isomorphic (see [4]). Therefore if R is reductive then A^H is affine for all such affine A iff $k[R/H]$ is affine. Moreover, preserving A^H one can assume H to be observable (i.e., R/H to be quas affine variety, cf. [5]). Let us call H a *Grosshans subgroup* of R if H is observable and the algebra $k[R/H]$ is affine. (It was in [5] where the role of such subgroups was clarified.) Presently one has to consider the original Hilbert's fourteenth problem as the problem of classification of Grosshans subgroups of $\text{GL}(V)$. Note that the classification of observable subgroups of reductive groups ($\text{char } k = 0$) was obtained by A. Sukhanov (1977).

It follows from the above that in the problem (b) of the classical scheme of "the explicit description" of $k[V]^G$ it is natural to assume that G is a reductive group. Also let $\text{char } k = 0$. The problem was solved in 1868–1870 by P. Gordan for invariants of systems of binary forms and in 1893 by Hilbert for invariants of systems of forms in arbitrary many variables—these are the heights of the classical theory. For each fixed d the space $k[V]_d^G$ of all homogeneous elements of degree d of $k[V]^G$ can be described explicitly using the methods of representation theory. This reduces (b) to the problem of explicitly finding such a number $M_{G,V}$ that $k[V]^G$ is generated by invariants of degree $\leq M_{G,V}$, cf. [6]. We know such $M_{G,V}$ now:

THEOREM 2. *Let $n = \dim V$, $s = \dim G$, $r = \text{rk } G$, let T be a maximal torus of G and $(k^*)^r \rightarrow T$ a fixed isomorphism given in coordinates by the formula*

$$(t_1, \dots, t_r) \mapsto (f_{ij}(t_1, \dots, t_r)), \quad f_{ij} \in k(t_1, \dots, t_r).$$

Let t be the greatest of numbers m_l taken over all $1 \leq l \leq r$ and all monomials $t_1^{m_1} \dots t_r^{m_r}$ that occur in at least one f_{ij} . For each $h > 0$ let $C(h) = \text{LCM}\{a \in \mathbb{N} \mid 0 < a \leq h\}$. Then one can take:

- (1) $M_{G,V} = |G|$ if G is finite;
- (2) $M_{G,V} = n \cdot C(n \cdot s! \cdot t^s)$ if $G = T$ is a torus;

(3)

$$M_{G,V} = n \cdot C \left(\frac{2^{r+s} n^{s+1} (n-1)^{s-r} t^r \cdot (s+1)!}{3^s ((s-r)/2!)^2} \right)$$

if G is semisimple and connected.

The proof of (1) was given by E. Noether [7], of (2) by G. Kempf [8], and of (3) by the author [6]. The case of an arbitrary reductive G is easily reduced to these three cases, cf. [8].

The estimates $M_{G,V}$ given by this theorem are large, and the problem remains to find its improvement. However, in any case a general solution to the problem (b) will be only of a theoretical value because nowadays it is precisely clear (see below) that the practical finding of generators of a “general” $k[V]^G$ is not a reasonable problem. Nevertheless, using some general results on actions one can distinguish the classes of actions $G: V$ admitting a deep investigation (right until the explicit description of $k[V]^G$). This shaped the point of view that the classes of actions $G: V$ distinguished by various “nice” properties are of paramount importance for invariant theory while the others $G: V$ are in some way “wild.” Conditionally, one can distinguish three sources of such results: algebraic geometry, commutative algebra, and representation theory.

2. Geometrical results on actions and the description of invariants.

Further let $\text{char } k = 0$ and G be a reductive group which acts on an affine variety X . We assume that all the actions are effective.

It was a long time ago when the idea appeared to simplify the problem of the description of $k[X]^G$ by finding a subgroup $H \subset G$ and an H -invariant subvariety $Y \subset X$ such that the restriction of functions defines an isomorphism $k[X]^G \rightarrow k[Y]^H$. One can always assume that $H = N_G(Y) = \{g \in G \mid g(Y) \subset Y\}$ because $k[Y]^H = k[Y]^{N_G(Y)/Z_G(Y)}$, where $Z_G(Y) = \{g \in G \mid g(y) = y \ \forall y \in Y\}$. Let us call Y a *Chevalley section*. (Apparently it was in the Chevalley restriction theorem, cf. [9], where such a Y was used explicitly for the first time.)

The systematic method to construct Chevalley sections was discovered by D. Luna and R. Richardson [9]. Let $X/G = \text{Specm } k[X]^G$ and $\pi_{G,X}: X \rightarrow X/G$ be a morphism induced by the embedding $k[X]^G \hookrightarrow k[X]$. It is G -invariant, surjective, and there is a unique closed orbit $O(\xi)$ in each fiber $\pi_{G,X}^{-1}(\xi)$ [2, 10]:

THEOREM 3. *Let X be irreducible and smooth. Then (a) [10] there exists an open set $\Omega \subset X/G$, $\Omega \neq \emptyset$, such that $\pi_{G,X}^{-1}(\xi)$ and $\pi_{G,X}^{-1}(\eta)$ (and hence $O(\xi)$ and $O(\eta)$) are G -isomorphic $\forall \xi, \eta \in \Omega$. Specifically, there exists a reductive subgroup $C \subset G$ such that for any $\xi \in \Omega$ and $v \in O(\xi)$ the stabilizer G_v of v is conjugate to C . (b) [9] There exists an irreducible component of the variety $X^C = \{x \in X \mid C(x) = x\}$ which is a Chevalley section for $G: X$.*

All of this is applicable to a linear action $G: V$ in which case $Y = V^C$ is a linear subspace and $N_G(Y) = N_G(C)$. One has to consider Y and $N_G(Y)/Z_G(Y)$ as the analogs of a Cartan subspace and a Weyl group in the general situation (they turn exactly into these objects in the case of adjoint action $G: \text{Lie } G$).

The reduction of $G: V$ to $N_G(C): Y$ is not a tautological one iff $C \neq \{e\}$. This explains the interest in a description of $G: V$ such that $C \neq \{e\}$. This description is based on the fact that under the conditions of Theorem 3 there exists "a generic stabilizer," i.e., a subgroup $G_* \subset G$ such that G_v is conjugate to G_* for all v from a nonempty open set in X , and moreover, $G_* \subset C$, see [30, 10]. It is easier to find G_* than C ; on the other hand, $G_* \neq \{e\}$ implies $C \neq \{e\}$. It is clear that $G_* = \{e\}$ if G is either finite or a torus. At present the classification (list) of $G: V$ such that $G_* \neq \{e\}$ is known if either (a) G is connected simple or (b) G is connected semisimple and $G: V$ is irreducible. A. Elashvili (1972) described such $G: V$ with $\dim G_* > 0$ and found $\text{Lie } G_*$ explicitly; A. Popov (1975–78) described such $G: V$ with finite G_* and found G_* explicitly. (They developed, resp., some ideas of the papers by E. Andreev, E. Vinberg, A. Elashvili (1967), and E. Andreev and the author [11].) It follows from [11] that for a fixed G there exists, under the conditions (a) and (b), only finitely many action $G: V$ with $V^G = \{0\}$ and $G_* \neq \{e\}$; this is no longer true for a connected semisimple G when (a) and (b) are infringed, though $G_* = \{e\}$ for "a general" $G: V$. (An early result of this sort was obtained by D. Hadziev (1967).)

$G_* \neq C$ in general. If $G_* = C$ then $G: X$ is said to be *stable* [12] (stability means closedness of "the generic" orbit). For a linear action $G: V$ a stability criterion was given by the author [12]: if G is connected semisimple (the general case reduces to this one [13]), then $G: V$ is stable iff G_* is reductive. This means that stable $G: V$ are "typical" and makes it possible to give, under the conditions (a) and (b) from above, an explicit classification (list) of nonstable $G: V$ —there turns out to be only a few of them.

Those $G: V$ with a finite "Weyl group" $N_G(Y)/Z_G(Y)$ are especially close to the model of adjoint action. For all such known $G: V$ the algebra $k[V]^G$, when described with ad hoc methods, turned out to be free. D. Panjushev [14] showed that these facts have a general origin:

THEOREM 4. *Let G be connected and for each subvariety $Z \subset V/G$ of codimension > 1 one has*

$$\text{codim } \pi_{G,V}^{-1}(Z) > 1$$

(this is automatically fulfilled if either G is semisimple, or G is a subgroup of orthogonal group $O(V)$, or $\pi_{G,V}$ is equidimensional). If $k[V]^G$ is isomorphic to an algebra of invariants of a finite linear group, then both of these algebras are free.

This means that the theories of invariants of connected and of finite linear groups cannot be reduced to each other and confirms the opinion that these are two "different" theories.

There exist examples of proper linear subspaces in V which are Chevalley sections but cannot be obtained by means of Theorem 3 ($C = \{e\}$). But another general result on existence of Chevalley sections besides Theorem 3 is not known. The fundamental role in the proof of Theorem 3 as well as of many other results on actions $G: X$ is played by Luna's étale slice theorem [10].

3. Commutative algebra and invariant theory. At present one can distinguish two approaches to a more delicate (than finding generators) description of a multiplicative structure of $k[V]^G$.

The first one is inspired by the following result of M. Hochster and J. Roberts [15]:

THEOREM 5. *If X is smooth, then $k[X]^G$ is a Cohen-Macaulay algebra.*

In fact J.-L. Boutot [16] later proved the more general

THEOREM 6. *If X has rational singularities, then X/G also has rational singularities.*

For a linear action $G: V$, Theorem 5 means the existence of such homogeneous elements $a_1, \dots, a_m, b_1, \dots, b_p \in k[V]^G$ that a_i (parameters) are algebraically independent and $k[V]^G = \bigoplus_{i=1}^p k[a_1, \dots, a_m]b_i$. This reduces the description of a structure of $k[V]^G$ to finding "a multiplication-table" for b_j and gives the formula for the Poincaré series:

$$P_{G,V}(t) = \sum_{d \geq 0} (\dim k[V]_d^G) t^d = \sum_{i=1}^p t^{\deg b_i} / \prod_{j=1}^m (1 - t^{\deg a_j}).$$

G. Kempf proved (1979) that $\deg P_{G,V}(t) \leq 0$; therefore it follows from here that $\max_i b_i \leq \sum_{j=1}^m \deg a_j$. This means that one can constructively find b_i whenever a_j are given. To constructively find a_j is a difficult problem. It is solved for a finite G by E. Dade, see [24]. (One has to choose $f_j \in V^*$, $1 \leq j \leq n$, as follows: $f_1 \neq 0$, $f_{j+1} \notin \langle g_1(f_1), \dots, g_j(f_j) \rangle \forall g_1, \dots, g_j \in G$; then a_j is a product of all elements of the orbit $G(f_j)$.) For a connected G no recipe of this sort is known, but there is an ideology, going back to Hilbert [31], which is based on the characterization of the parameters as algebraically independent elements a_j such that $\{v \in V \mid a_j(v) = 0 \forall j\}$ coincides with the "null-cone" $\pi_{G,V}^{-1}(\pi_{G,V}(0))$ and on the geometrical description of the "null-cone" delivered by the Hilbert-Mumford theorem [2]. One obtains in this way explicit upper estimates of $\deg a_j$ as well as the proof of (2) and (3) in Theorem 2, see [8, 6].

The second approach is given by the classical theory of syzygies. Let y_1, \dots, y_d be a minimal system of homogeneous generators of $k[V]^G$ (here d is the *embedding dimension* $\text{ed } k[V]^G$ of $k[V]^G$), $A = k[x_1, \dots, x_d]$ the polynomial k -algebra in indeterminates x_i with the grading given by $\deg x_i = \deg y_i$, and I the kernel of the epimorphism $A \rightarrow k[V]^G$, $x_i \mapsto y_i$. I is the first module of syzygies of $k[V]^G$; its elements are the relations (syzygies) which hold among the y_i . Taking a minimal system of homogeneous generators of A -module I , one can present I as the image of a free A -module M_1 under a homogeneous epimorphism of A -modules $\varphi_1: M_1 \rightarrow I$; then do the same with $\text{Ker } \varphi_1$, and so on. One obtains then a minimal free resolution of A -module I which is finite by the Hilbert syzygy theorem:

$$0 \rightarrow M_h \xrightarrow{\varphi_h} \dots \xrightarrow{\varphi_2} M_1 \xrightarrow{\varphi_1} I \xrightarrow{\varphi_0} 0.$$

Its length h is the *homological dimension* $\text{hd } k[V]^G$ of $k[V]^G$, and $\text{Ker } \varphi_{i-1}$ is the i th *module of syzygies*. One has $h = d - \dim V/G$ because of Theorem 5.

A complete description of all $\text{Ker } \varphi_i$ has been obtained only in a number of special cases. In the nineteenth century an answer was found only for invariants of binary forms of degree $v \leq 6$ (and also for several other elementary cases), and the case $v = 8$ was understood only comparatively recently by T. Shioda (1967). Classics described I for the actions of classical groups on systems of vectors and covectors, but the higher syzygies for them were described only by A. Lascoux (1978) and, recently, by some others. The reason for this is the complexity of the structure of $k[V]^G$ in the "general case." This was definitely known to classics, but the precise meaning of this assertion was elucidated only recently.

The natural measure of complexity of $k[V]^G$ is the number $h = \text{hd } k[V]^G$. One can take h as the base for the systematic investigation of algebras $k[V]^G$ of different actions $G: V$ of a given fixed G : instead of $G: V$ taken by chance one has first of all to investigate those $G: V$ for which h is small. This program was put on firm ground by the following theorem proved by the author [13, 23]:

THEOREM 7. *Let G be either connected semisimple or finite. Then for any $h \geq 0$ there exist, to within isomorphism and addition of a trivial direct summand, only finitely many finite-dimensional rational G -modules V such that $\text{hd } k[V]^G = h$.*

It follows from here that for $G \neq \{e\}$ there exists such a G -module V that $\text{hd } k[V]^G > h$ (and the set of such V is infinite). This gives a precise meaning to the assertion about complexity of $k[V]^G$ in "the general case." The estimates from [13] give quantitative expression to this. E.g., if $G = \text{SL}_2$ and V is the module S_n of binary forms of degree n , then for any prime p and $n > 4p + 5$ one has

$$\text{hd } k[S_n]^G \geq (p-1) \binom{[n/p] + p - 2}{p} + \frac{p-1}{2} ([n/p] - 1)^2 + [n/p] - n + 2$$

and therefore $\text{hd } k[V]^G$ grows faster than any power of n as $n \rightarrow \infty$. If G is an exceptional simple group and V is a nontrivial irreducible G -module different from adjoint, elementary or its dual, then $\text{hd } k[V]^G \geq 9208, 26335, 3403, 2278$ and 3 resp. for $G = E_8, E_7, E_6, F_4$ and G_2 . On the other hand, Theorem 7 and the constructive character of its proof show that the state of affairs turns out to be not so hopeless as might first be imagined and make natural the problem of classifying, for a given group G and integer h , all $G: V$ with $\text{hd } k[V]^G = h$. It is natural to do this sequentially, from smaller values of h to larger, since the method itself contains the possibility of induction: it is based, specifically, on the monotonicity theorems, according to which ed and hd do not increase under passage from a representation to a subrepresentation or a slice representation. These reflect a general "principle of inheritance," according to which the properties of $k[V]^G$ are "not made worse" (inherited) under indicated passages; there are a lot of other evidences in favor of it.

The first case of this program is the description of such $G: V$ that $\text{hd } k[V]^G = 0$, i.e. $k[V]^G$ is free (*property (FI)*). If G is finite, then the answer is given by

THEOREM 8. *Let G be a finite subgroup of $\text{GL}(V)$. Then the following properties are equivalent: (a) G is generated by pseudoreflections; (b) $k[V]^G$ is free.*

Proof was given by G. Shephard and J. Todd (1954) who also gave a classification (list) of all G with the property (a), and C. Chevalley (1955) (for (a) \Rightarrow (b)).

For a connected G the method of classification of $G: V$ with a free $k[V]^G$ was discovered only comparatively recently by V. Kac, the author, and E. Vinberg [25], who also tested it on the case of connected simple G and irreducible $G: V$ (it is based on "the principle of inheritance" of the property (FI):

THEOREM 9. *Let G be a connected simple linear irreducible group. Then the following properties are equivalent: (1) G has a free algebra of invariants; (2) G is one of the following groups: (a) Case when $k[V]^G = k$: SL_n, Sp_n (n is even), $\Lambda^2 \text{SL}_n$ (n is odd), Spin_{10} ; (b) Case when $\dim V/G = 1$: SO_n ($n \neq 4$), $\Lambda^2 \text{SL}_n$ (n is even), $S^2 \text{SL}_n, \Lambda^3 \text{SL}_n$ ($n = 6, 7, 8$), $\Lambda_0^3 \text{Sp}_6, S^3 \text{SL}_2, \text{Spin}_n$ ($n = 7, 9, 11, 12, 14$), E_6, E_7, G_2 ; (c) Adjoint groups of simple Lie algebras; (d) Isotropy groups of symmetric spaces: $\Lambda_0^2 \text{Sp}_n$ (n is even), $S_0^2 \text{SO}_n, \Lambda^4 \text{SL}_8, \Lambda_0^4 \text{Sp}_8, \text{Spin}_{16}, F_4$; (e) The others: $\Lambda^3 \text{SL}_9, S^3 \text{SL}_3, \text{Spin}_{13}$. (Here $\Lambda_0^m G$ and $S_0^m G$ is the irreducible component of highest weight resp. of $\Lambda^m G$ and $S^m G$.)*

Using this method, G. Schwarz (1978) and independently O. Adamovich and E. Golovina (1979) gave the list of reducible actions $G: V$ of connected simple groups G with a free algebra of invariants.

It is unknown whether it is possible to describe $G: V$ with a connected G and free $k[V]^G$ by means of a general group-theoretical characterization (as in a theorem like Theorem 8) instead of giving the list. (A naïve attempt to carry (a) \Rightarrow (b) of Theorem 8 to infinite G turned out to be unsound as A. Zalesskii showed (1983).) However, several general methods to construct $G: V$ with a free $k[V]^G$ are known at present. The first one is E. Vinberg's Θ -groups [26]. These are defined with an aid of semisimple automorphism Θ of a connected reductive group H as linear groups defined by the adjoint action of $(H^\Theta)^0$ on the eigenspaces of $d\Theta$ in $\text{Lie } H$; for $\Theta = e$ one obtains the adjoint groups and for $\Theta^2 = e$, $\Theta \neq e$ the isotropy groups of symmetric spaces. Therefore the Chevalley restriction theorem as well as classical results of B. Kostant [18] and B. Kostant and S. Rallis (1971) follows from the theory [26]. These groups are included into a wider class of groups with a free $k[V]^G$, polar groups of J. Dadok and V. Kac [27]. (They are defined by the condition that for some closed orbit Gv one has

$$\dim\{x \in V \mid (\text{Lie } G)x \subseteq (\text{Lie } G)v\} = \dim V/G.)$$

The polarity *property (P)* is inheritable; besides a general theory, a classification (list) of all connected simple polar linear G is given in [27]. All the groups listed in Theorem 9 are polar and almost all are Θ -groups or their commutator

subgroups (namely, with 2 exceptions). The third method was discovered by D. Panjushev (1986): these are the isotropy groups of homogeneous spaces G/H where $G \supset H$ are connected reductive groups and (G, H) is a spherical pair (i.e., a Borel subgroup B of G has an open orbit in G/H); such pairs were classified in 1985 by I. Mikitjuk and independently and with another method by M. Brion. Polar groups have a (linear) Chevalley section with a finite “Weyl group” and the property (FI) for them follows from Theorem 4.

The next case is $\text{hd } k[V]^G = 1$, i.e., V/G is a hypersurface. Hence (as well as for $\text{hd } k[V]^G = 2$ if G is either connected semisimple or finite and lies in $\text{SL}(V)$), $k[V]^G$ is a complete intersection (i.e., $\text{rk } M_1 = \text{ed } k[V]^G - \dim V/G$). This property is inheritable. V. Kac, K.-I. Watanabe, and independently N. Gordeev (1982) found an analog of (b) \Rightarrow (a) of Theorem 8:

THEOREM 10. *If G is a finite subgroup of $\text{GL}(V)$ and $k[V]^G$ is a complete intersection, then G is generated by the set $\{g \in G \mid \text{rk}(g - 1) \leq 2\}$.*

The proof is ideologically close to the proof of Theorem 4 and based on simple connectivity of a certain open subset of V/G . (It is worth noting here that, as M. Mikhailova and O. Schwarzman proved (1985), a finite group $H \subset \text{GL}(\mathbf{R}^n)$ is generated by the set $\{g \in H \mid \text{rk}(g - 1) = 2\}$ iff \mathbf{R}^n/H is homeomorphic to \mathbf{R}^n .) This supplies a great deal of information and at present we know the complete classification (list) of finite $G \subset \text{GL}(V)$ such that $k[V]^G$ is a complete intersection (N. Gordeev and H. Nakajima (1985–86)) and, particularly, hypersurface (H. Nakajima (1983)). H. Nakajima (1984) found all connected simple irreducible linear G with the same property; earlier N. Beklemishev (1983) and V. Kac (1983) found such G of the type $S^d \text{SL}_n$.

The cases of a bigger $\text{hd } k[V]^G$ were as yet considered only in the classical situation of invariants of binary forms: the following was proved by the author.

THEOREM 11 [17]. (1) $\text{hd } k[S_n]^{\text{SL}_2} \leq 10$ iff $n = 0, 1, 2, 3, 4, 5, 6$, and 8 (and $\text{hd } k[S_n]^{\text{SL}_2} \geq 18, 34, 28$ for $n = 7, 10, 12$ resp.). (2) $\text{hd } k[V]^{\text{SL}_2} = h \leq 3$ and $V^{\text{SL}_2} = \{0\}$ iff: (a) $h = 0$: $V = S_1, S_2, S_3, S_4, 2S_1, S_1 \oplus S_2, 2S_2, 3S_1$; (b) $h = 1$: $V = S_5, S_6, S_2 \oplus S_3, S_2 \oplus S_4, 2S_1 \oplus S_2, S_1 \oplus 2S_2, 3S_2, S_1 \oplus S_3, S_1 \oplus S_4, 2S_4, 4S_1$; (c) $h = 2$: $V = 2S_3$; (d) $h = 3$: $V = S_8, 5S_1$.

This shows that the classics (who investigated such cases that turned out to be good to work on) did not miss any sufficiently simple case.

Therefore, considering a hierarchy of “nice” properties of algebras: (free algebra) \Rightarrow (hypersurface) \Rightarrow (complete intersection) \Rightarrow (Gorenstein) \Rightarrow (Cohen-Macaulay), we have at present a satisfactory description of $k[V]^G$ with such properties. ($k[V]^G$ is Gorenstein if either G is connected semisimple (this follows from [15]) or finite and lies in $\text{SL}(V)$, and moreover, on the contrary, if G is finite and does not contain pseudoreflections and $k[V]^G$ is Gorenstein, then $G \subset \text{SL}(V)$, K. Watanabe (1974), cf. [24].) It was proved by R. Stanley (1978) that $k[V]^G$ is Gorenstein iff

$$P_{G,V}(t^{-1}) = (-1)^{\dim V/G} t^{-\deg P_{G,V}} P_{G,V}(t)$$

(functional equation); this is used [23] in the proof of Theorem 7 for $h = 0$.

4. Invariant theory and representation theory. One can consider $f \in k[V]^G$ as an equivariant morphism $f: X \rightarrow k$. Changing k to any G -module W one obtains a more general notion—a covariant. Its existence is equivalent to an existence of a homomorphism $W^* \rightarrow k[X]$ of G -modules. Hence description of covariants reduces to description of G -module structure of $k[X]$. This is convenient to do by considering $k[X]$ as a $k[X]^G$ -module (module of covariants): if \mathfrak{S} is the set of equivalence classes of simple rational G -modules and $k[X]_\lambda$ is an isotypical component of $k[X]$ of type $\lambda \in \mathfrak{S}$, then $k[X]$ is a $k[X]^G$ -module of finite type and $k[X] = \bigoplus_{\lambda \in \mathfrak{S}} k[X]_\lambda$. If G is connected, then one can identify \mathfrak{S} with the monoid of dominant weights with respect to B and the highest vectors in $k[X]$ form so-called *algebra of covariants* $k[X]^U$ (U is the unipotent radical of B). In the classical situation of a linear action $G: V$ the case of a free module of covariants (*property (FM)*) is particularly interesting. In this case $k[V] = H \otimes k[V]^G$, where H is a G -invariant complement in $k[V]$ to the ideal generated by $\bigoplus_{i \geq 1} k[V]_i^G$, and $k[V]_\lambda$ is a free $k[V]^G$ -module of the rank m_λ which is equal to the multiplicity of a G -module W of the type λ in H . G. Schwarz showed (1978) that m_λ can be found by means of C (see Theorem 3): V is a sum of a trivial C -module, the isotropy module of G/C , and a C -module L with $L^C = \{0\}$, and $m_\lambda = \dim(k[L] \otimes W^*)^C$. The answer is particularly simple when $C = G_*$, in which case $L = \{0\}$ and $m_\lambda = \dim(W^*)^C$.

The problem of description of those $G: V$ with the property (FM) was for the first time considered by B. Kostant [18] who investigated adjoint action. The general investigation was started by the author [19]; it is based on

THEOREM 12. *The property (FM) for $G: V$ is equivalent to both of the properties together: (a) $k[V]^G$ is free; (b) $\pi_{G,V}$ is equidimensional.*

Therefore if G is finite then (FM) iff G is generated by pseudoreflections (and in this case $G: H$ is a regular representation). If G is connected, then the equidimensionality of $\pi_{G,V}$ is an independent condition. This naturally leads to an independent problem of classification of those $G: V$ with the equidimensional $\pi_{G,V}$ (*property (E)*). “Null-cone” $\pi_{G,V}^{-1}(\pi_{G,V}(0))$ is the fiber of a maximal dimension, and one can derive estimates of its dimension using its geometrical description (see, e.g., [19]). Besides this, (FM) and (E) are inheritable properties. One obtains in this way the following

THEOREM 13. *Let G be either connected semisimple or finite. Then there exist, to within isomorphism and addition of a trivial direct summand, only finitely many finite-dimensional rational G -modules V with the property (E) (and hence with the property (FM)).*

As a matter of fact one can give a quantitative expression to this: e.g., if G is connected semisimple and $V^G = \{0\}$, then (E) implies $\dim V \leq 3 \dim G - 3$. Moreover, it is possible to give the complete classification in many cases: the author found the lists of all irreducible $G: V$ with a connected simple G which

has properties resp. (E) and (FM) [19]; afterwards these were extended to the cases of reducible $G: V$ with a connected simple G by O. Adamovich (1980) and G. Schwarz (1978) and irreducible $G: V$ with a connected semisimple G by P. Littelmann (1986). The general group-theoretical characterization (instead of the list) of those $G: V$ with the properties (E) or (FM) is unknown, but there are several general methods to construct such $G: V$: Θ -groups, polar groups, isotropy groups of reductive spherical pairs have properties (E) and (FM). (E) is a priori fulfilled for so-called *visible actions* (*property* (FO)), i.e., such that each fiber of $\pi_{G,V}$ contains only finitely many orbits; V. Kac (1975) gave a list of all irreducible visible $G: V$ with connected G . Taking an adjoint action as a model, one can also consider two other “nice” properties: (SC)—some linear subspace of V is a Chevalley section with a finite “Weyl group”; (SW)—some linear subvariety of V (“Weierstrass section”) intersects each fiber of $\pi_{G,V}$ in a unique point. A priori: $(SC) \Rightarrow (FM) \Leftarrow (SW)$. The comparison of the lists shows the following remarkable coincidence:

THEOREM 14. *Let G be connected simple and $G: V$ irreducible. Then the following properties are equivalent: (1)(FI); (2)(FM); (3)(E); (4)(FO); (5)(SC); (6)(SW); (7) $G_* \neq \{e\}$; (8)(P); (9) G is a group from the list of Theorem 9.*

This cannot be extended to a general case (G is connected semisimple) though there are different implications between these properties.

Besides these, there is at least one other “nice” property: (FC)—algebra of covariants is free. A priori: $(FC) \Rightarrow (FM)$. M. Brion (1985) found the list of all $G: V$ with connected simple G which have property (FC). D. Panjushev (1985) discovered for $k[X]^U$ an analog of construction from Theorem 3 which gives “a Chevalley section.” It is generally true that $k[X]$ and $k[X]^U$ have some important properties in common (e.g., rational singularities), cf. [4]. One can also “restore” to a certain extent $k[X]$ from $k[X]^U$ which permits one to present any action $G: X$ as a flat deformation (with one-dimensional base) of a certain very special action, cf. [4, 20]. In its turn this permits one to reduce some problems about $G: X$ to the case of special actions, see [20].

The fact that only a few of $G: V$ have “nice” properties makes one hope that one can effectively solve for such $G: V$ the main problem of invariant theory: the classification of orbits and the description of $k[V]^G$. This actually turns out to be possible to do: besides the cases considered by the classics, at present we know the solution, e.g., for Spin_n , $n \leq 12$ (J.-I. Igusa (1970)), $n = 13$ (V. Gatti, E. Viniberghi (1978)), $n = 14$ (the author (1978)), $n = 16$ (L. Antonjan, A. Elashvili (1982)), $\Lambda^d \text{SL}_n$, $(d, n) = (3, 9)$ (E. Vinberg, A. Elashvili (1978)), $(d, n) = (4, 8)$ (L. Antonjan (1980)), $S^3 \text{SL}_4$ (N. Beklemishev (1981–1982)).

5. Birational invariant theory. This is the other side of the classical theme which concerns the description of the field $k(V)^G$ of invariant rational functions on V for a linear $G: V$. The main problem here (posed in fact by E. Noether (1913)) is whether $k(V)^G$ is rational or not. Important progress

was made recently. D. Saltman [21] gave an example of a finite G such that $k(V)^G$ is not stably rational (hence is not rational). He based his example on the consideration of a certain subgroup Br_0 of $Br(k(V)^G)$ which has to be trivial if $k(V)^G$ is stably rational (and which is not trivial in his example). It turns out that Br_0 depends only on G but not on $G:V$. (This was in essence proved by D. Faddeev (1951) because all $k(V)^G$ are stably isomorphic for different $G:V$.) F. Bogomolov (1986) proved that $Br_0 = \{\alpha \in H^2(G, \mathbf{Q}/\mathbf{Z}) \mid \alpha|_A = 0 \text{ for each abelian subgroup } A \subset G\}$. There are no examples of nonrational $k(V)^G$ for connected G up to now. E. Vinberg (1982) proved that $k(V)^G$ is rational if G is connected solvable. F. Bogomolov (1986) proved stable rationality of $k(V)^G$ in the case when G is a connected simply connected reductive group without factors of the type B_n, D_n, E_8 and $G_* = \{e\}$. An important innovation was made by P. Katsilo [22] who introduced a suitable analog of the Chevalley section, namely, an irreducible $Y \subset V$ such that $\overline{GY} = V$ and $g(y) \in Y \forall g \in G, y \in Y$, implies $g \in N_G(Y)$. The restriction defines an isomorphism $k(V)^G \rightarrow k(Y)^{N_G(Y)}$; another important property is that one can “reproduce” such Y by means of covariants $W \rightarrow V$ (“lifting”). Using such Y , P. Katsilo (1984) proved the rationality of $k(V)^G$ for almost all actions of $G = \mathrm{SL}_2$ and $G = \mathrm{SL}_2 \times k^*$ (k^* acts by homotheties) and hence the rationality of the moduli space of hyperelliptic curves of genus g (for $g = 4$ one needs some extra considerations). It is worth noting that the problem is closely related to the problem of rationality of homogeneous spaces: the stable rationality of $k(V)^G$ for $G:V$ with $G_* = \{e\}$ is equivalent to the stable rationality of GL_n/G for any embedding $G \subset \mathrm{GL}_n$. (Hence there exists a nonrational G/H with a connected reductive G and finite H ; it is unknown whether it is possible for a connected reductive H .)

6. Conjectures and problems. In conclusion, I want to take this opportunity to point out several conjectures and problems which seem to me to be of some interest.

1. CONJECTURE. $(E) \Rightarrow (FM)$ for a connected G . This was for the first time formulated by the author in [19] and since then got some popularity as the “Russian conjecture.” Apparently, a general fact of commutative algebra is hidden beyond this. The conjecture is a posteriori proved now (by the inspection of the lists) for simple G and for irreducible $G:V$ of semisimple G .

2. CONJECTURE. Let G be connected and semisimple. Then $\deg P_{G,V} + \dim V \geq 0$ for each $G:V$.¹ It follows from here that if $k[V]^G$ is free and $V^G = \{0\}$ then $\dim V \leq 2 \dim G$, and if moreover $G:V$ is irreducible then $\dim V \leq (3 \dim G + 1)/2$. Hence this gives another approach to the classification of $G:V$ with the property (FI). It is proved in [23] that in fact $\deg P_{G,V} + \dim V = 0$ for almost all of $G:V$ if either G is simple or $G:V$ is irreducible.

3. CONJECTURE. Let H be a unipotent subgroup of G such that $N_G(H)$ contains a maximal torus of G . Then H is a Grosshans subgroup. Is it possible

¹Right before my departure from Moscow to Berkeley I received the preprint of [32] where the proof of this conjecture is given.

to find a subgroup $S \subset G$ such that $\{\lambda \in \mathfrak{S} \mid k[G/S]_\lambda \neq 0\}$ is not finitely generated?

4. Classify $G:V$ with: (a) a nontrivial linear Chevalley section; (b) the property (SC); (c) the property (SW); (d) the property that $\pi_{G,V}^{-1}(\xi) \forall \xi \in V/G$ is reduced.

5. Let $G \supset H$ be a spherical pair and H not reductive. Is the algebra of invariants of the isotropy representation free? Classify all such pairs $G \supset H$.

REFERENCES

1. M. Nagata, *Lectures on the fourteenth problem of Hilbert*, Tata Inst. Fund. Res. Lectures on Math. and Phys., vol. 31, Tata Inst. Fund. Res., Bombay, 1965.
2. D. Mumford and J. Fogarty, *Geometric invariant theory*, 2nd ed., Ergeb. Math. Grenzgeb. (3), vol. 34, Springer-Verlag, 1982.
3. V. L. Popov, *Hilbert's theorem on invariants*, Dokl. Akad. Nauk SSSR **249** (1979), 551–555= Soviet Math. Dokl. **20** (1979), 1318–1322.
4. —, *Contraction of the actions of reductive algebraic groups*, Mat. Sb. **130**(172) (1986), 310–334= Math. USSR Sb. **58** (1987) (to appear).
5. F. Grosshans, *Observable groups and Hilbert's fourteenth problem*, Amer. J. Math. **95** (1973), 229–253.
6. V. L. Popov, *The constructive theory of invariants*, Izv. Akad. Nauk SSSR Ser. Mat. **45** (1981), 1100–1120= Math. USSR Izv. **19** (1982), 359–376.
7. E. Noether, *Der Endlichkeitssatz der Invarianten endlicher Gruppen*, Math. Ann. **77** (1916), 89–92.
8. G. R. Kempf, *Computing invariants*, Preprint, 1986.
9. D. Luna and R. W. Richardson, *A generalization of the Chevalley restriction theorem*, Duke Math. J. **46** (1979), 487–496.
10. D. Luna, *Slices étales*, Bull. Soc. Math. France **33** (1973), 81–105.
11. E. M. Andreev and V. L. Popov, *The stationary subgroups of points in general position in a representation space of a semisimple Lie group*, Funktsional. Anal. i Prilozhen. **5** (1971), no. 4, 1–8= Functional Anal. Appl. **5** (1971), 265–271.
12. V. L. Popov, *Stability criteria for the action of a semisimple group on a factorial manifold*, Izv. Akad. Nauk SSSR Ser. Mat. **34** (1970), 523–531= Math. USSR Izv. **4** (1970), 527–535.
13. —, *Syzygies in the theory of invariants*, Izv. Akad. Nauk SSSR Ser. Mat. **47** (1983), 544–622= Math. USSR Izv. **22** (1984), 507–585.
14. D. I. Panyushev, *On orbit spaces of finite and connected linear groups*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), 95–99= Math. USSR Izv. **20** (1983), 97–101.
15. M. Hochster and J. Roberts, *Rings of invariants of reductive groups acting on regular rings are Cohen–Macaulay*, Adv. in Math. **13** (1974), 125–175.
16. J.-F. Boutot, *Singularités rationnelles et quotients par les groupes réductifs*, Invent. Math. **88** (1987), 65–68.
17. V. L. Popov, *Homological dimension of algebras of invariants*, J. Reine Angew. Math. **341** (1983), 157–173.
18. B. Kostant, *Lie group representations in polynomial rings*, Amer. J. Math. **85** (1963), 327–404.
19. V. L. Popov, *Representations with a free module of covariants*, Funktsional. Anal. i Prilozhen. **10** (1976), no. 3, 91–92= Functional Anal. Appl. **10** (1976), 242–243.
20. È. B. Vinberg, *Complexity of actions of reductive groups*, Funktsional. Anal. i Prilozhen. **20** (1986), no. 1, 1–13= Functional Anal. Appl. **20** (1986), 1–11.
21. D. Saltman, *Noether's problem over an algebraically closed field*, Invent. Math. **77** (1984), 71–84.
22. P. I. Katsylo, *The rationality of moduli spaces of hyperelliptic curves*, Izv. Akad. Nauk SSSR Ser. Mat. **48** (1984), 705–710= Math. USSR Izv. **25** (1985), 45–50.

23. V. L. Popov, *A finiteness theorem for representations with a free algebra of invariants*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), 347–370=Math. USSR Izv. **20** (1983), 333–354.
24. R. Stanley, *Invariants of finite groups and their applications to combinatorics*, Bull. Amer. Math. Soc. (N.S.) **1** (1979), 475–511.
25. V. G. Kac, V. L. Popov, and E. B. Vinberg, *Sur les groupes linéaires algébriques dont l'algèbre des invariants est libre*, C.R. Acad. Sci. Paris Sér. A Math. **283** (1976), 875–878.
26. È. B. Vinberg, *The Weyl group of a graded Lie algebra*, Izv. Akad. Nauk SSSR Ser. Mat. **40** (1976), 488–526=Math. USSR Izv. **10** (1976), 463–495.
27. J. Dadok and V. G. Kac, *Polar representations*, J. Algebra **92** (1985), 504–524.
28. M. Nagata, *Invariants of a group in an affine ring*, J. Math. Kyoto Univ. **3** (1964), 369–377.
29. W. Haboush, *Reductive groups are geometrically reductive*, Ann. of Math. (2) **102** (1975), 67–83.
30. R. W. Richardson, *Principal orbit types for algebraic transformation spaces in characteristic zero*, Invent. Math. **16** (1972), 6–14.
31. D. Hilbert, *Über die vollen Invariantensysteme*, Math. Ann. **42** (1893), 313–373.
32. F. Knop, *Über die Glattheit von Quotientenabbildungen*, Manuscripta Math. **56** (1986), 419–427.

MOSCOW INSTITUTE FOR ELECTRONIC MECHANICAL ENGINEERING, 109028 MOSCOW, USSR

Simple Lie Algebras over Fields of Prime Characteristic

ROBERT LEE WILSON

Dedicated to the memory of Boris Weisfeiler

The classification problem for finite-dimensional simple Lie algebras over a field F of prime characteristic p has attracted attention for about fifty years. Although it is still open, much progress has been made towards its solution. In particular, the known simple Lie algebras have a natural description (as discussed in Kostrikin's talk at the 1970 International Congress of Mathematicians [25]), the classification has been completed recently in several important special cases, and there are good techniques available for attacking the general problem.

We will assume throughout this article that F is algebraically closed of characteristic $p > 7$. For $p = 2, 3, 5$ counterexamples exist to some of the results stated here, and for $p = 7$ some of the proofs fail.

For an important class of Lie algebras over F , the restricted Lie algebras, the classification is now complete (see Theorem 3.3). We will (in §1) define these algebras and discuss some of the technical properties which make their study easier than that of general algebras. We will then describe the known finite-dimensional simple Lie algebras over F and discuss some of their structural properties (§2) and state three recent classification results (§3). We conclude (§4) by describing the main techniques used to establish these results: filtration theory and local analysis.

1. Restricted algebras. We will devote particular attention to *restricted* Lie algebras (= Lie p -algebras) over F (defined by Jacobson [14]). A Lie algebra L over F is restricted if it has a mapping $x \mapsto x^{[p]}$ satisfying

$$(1.1) \quad (\operatorname{ad} x)^p = \operatorname{ad}(x^{[p]}) \text{ for all } x \in L,$$

$$(1.2) \quad (\alpha x)^{[p]} = \alpha^p x^{[p]} \text{ for all } \alpha \in F, x \in L, \text{ and}$$

$$(1.3) \quad (x + y)^{[p]} = x^{[p]} + y^{[p]} + \sum_{i=1}^{p-1} s_i(x, y) \text{ where } s_i(x, y) \text{ is the coefficient of } t^{i-1} \text{ in } (\operatorname{ad}(tx + y))^{p-1}x.$$

If A is an associative algebra over F , the ordinary p th power map gives the Lie algebra A^- (with product $[x, y] = xy - yx$) the structure of a restricted Lie

Supported, in part, by NSF Grant #DMS-8603151.

algebra. Then any Lie subalgebra of A^- closed under taking p th powers is again a restricted Lie algebra. Applying this remark to $\text{Der } B \subseteq (\text{End } B)^-$ (where B is any algebra over F) shows that any derivation algebra is restricted.

Restricted Lie algebras are more tractable than arbitrary Lie algebras over F for a number of technical reasons. We note some of these. If L is a restricted Lie algebra, an element $x \in L$ is said to be *semisimple* if $x \in \text{span}\{x^{[p]}, x^{[p^2]}, \dots\}$ and is said to be *nilpotent* if $x^{[p^n]} = 0$ for some n . A *torus* T is a subalgebra (necessarily abelian) such that every $x \in T$ is semisimple. A subalgebra I is said to be *nil* if every $x \in I$ is nilpotent. The following theorems are due to Seligman [32].

THEOREM 1.1 (*Jordan-Chevalley-Seligman Decomposition*). *Let L be a restricted Lie algebra over F and $x \in L$. Then there exist unique elements $x_s, x_n \in L$ satisfying:*

- (a) $x = x_s + x_n$,
- (b) x_s is semisimple and x_n is nilpotent,
- (c) $[x_s, x_n] = 0$.

THEOREM 1.2. *Let L be a restricted Lie algebra over F . Then H is a Cartan subalgebra of L if and only if $H = C_L(T)$ ($= \{x \in L \mid [x, T] = (0)\}$) where T is a maximal torus in L .*

The following lemma (proved by applying the Engel-Jacobson theorem [16, Theorem 2.1] to $\bigcup_{\gamma \in T^*} I_\gamma$ and applying Theorem 1.1) will be useful.

LEMMA 1.3. *Let $(0) \neq I$ be a restricted ideal in a restricted semisimple Lie algebra L . Let T be a maximal torus in L . Then $I \cap T \neq (0)$.*

2. Description of the known simple Lie algebras over F . The known finite-dimensional simple Lie algebras over F are of two types: classical type (analogues over F of finite-dimensional simple Lie algebras over \mathbb{C}) and Cartan type (finite-dimensional analogues over F of the infinite Lie algebras of Cartan [9, 24, 34] over \mathbb{C}).

The classical algebras may be constructed as follows: Let L be a finite-dimensional simple Lie algebra over \mathbb{C} and let $L_{\mathbb{Z}}$ be a Chevalley lattice in L . Let $L_F = L_{\mathbb{Z}} \otimes_{\mathbb{Z}} F$. Then $L_F/\text{center } L_F$ is simple (and $\dim \text{center } L_F \leq 1$). Such algebras over F are called classical. Note that the algebras of types E-G are considered to be classical according to this definition.

If L is a classical simple Lie algebra over F then L is restricted and

- (2.1) L contains a Cartan subalgebra H which is a torus.

Letting Γ denote the set of roots of L with respect to H (so $L = H + \sum_{\gamma \in \Gamma} L_\gamma$, $L_\gamma = \{x \in L \mid [h, x] = \gamma(h)x \text{ for all } h \in H\}$) we have:

- (2.2) $\dim[L_\gamma, L_{-\gamma}] \leq 1$ for all $\gamma \in \Gamma$;
- (2.3) if $\alpha, \beta \in \Gamma$ then there is some $k \in \mathbb{Z}$ such that $\beta + k\alpha \notin \Gamma \cup \{0\}$;
- (2.4) if H_1, H_2 are Cartan subalgebras of L then there exists $\sigma \in \text{Aut } L$ such that $\sigma H_1 = H_2$.

In fact, (2.1)–(2.3) characterize the classical simple algebras.

THEOREM 2.1 (MILLS-SELIGMAN [29]). *Let L be a finite-dimensional simple Lie algebra over F satisfying (2.1)–(2.3). Then L is classical.*

We will say that a Lie algebra L over F is *compositionally classical* if all of its composition factors are classical or abelian.

Before giving the general definition of the algebras of Cartan type we will give a brief historical survey of the nonclassical simple Lie algebras.

The first nonclassical simple Lie algebra over F was discovered by Witt (see [44, 11]). We denote it by $W(1 : 1)$. It is defined as the algebra with basis $\{e_i | i \in \mathbb{Z}/(p)\}$ and multiplication

$$[e_i, e_j] = (j - i)e_{i+j}. \quad (2.5)$$

This algebra is restricted. Zassenhaus [44] obtained nonrestricted generalizations by letting G be a finite additive subgroup of F and defining multiplication on the algebra with basis $\{e_i | i \in G\}$ by (2.5). Further constructions of simple Lie algebras in this spirit (by giving an explicit multiplication table for a basis indexed, at least in part, by an additive group) were given by Kaplansky [22], Albert and Frank [1], and Block [4].

The Witt algebra $W(1 : 1)$ can be constructed in another way (due, independently, to Jacobson and Witt). Let $B_1 = F[x_1]/(x_1^p)$. Then letting D_1 denote the derivation of B_1 induced by $\partial/\partial x_1$, we see that $\text{Der } B_1$ has basis $\{x_1^i D_1 | 0 \leq i \leq p-1\}$ and that $e_i \mapsto (x_1 + 1)^{i+1} D_1$ gives an isomorphism of $W(1 : 1)$ onto $\text{Der } B_1$. This construction of $W(1 : 1)$ makes the following points apparent.

(2.6) $\text{span}\{x_1^i D_1 | i \geq 1\} (= W(1 : 1)_0)$ is a subalgebra of codimension one in $W(1 : 1)$. This subalgebra may be characterized as the unique compositionally classical subalgebra of maximal dimension in $W(1 : 1)$. It is therefore invariant under $\text{Aut } W(1 : 1)$.

(2.7) $x_1 D_1$ and $(x_1 + 1)D_1$ span maximal tori (which are also Cartan subalgebras) in $W(1 : 1)$. Since $x_1 D_1 \in W(1 : 1)_0$, $(x_1 + 1)D_1 \notin W(1 : 1)_0$ we see that these Cartan subalgebras are not conjugate under $\text{Aut } W(1 : 1)$. Thus (2.4) fails for $W(1 : 1)$.

(2.8) If $E^x = \sum_{i=0}^{p-1} (\text{ad } x)^i / i!$ then $E^{D_1}(x_1 D_1) = (x_1 + 1)D_1$. This, of course, implies that $E^{D_1} \notin \text{Aut } W(1 : 1)$. Call E^x a *Winter exponential* [42].

Jacobson [15] studied the algebras

$$W(m : 1) = \text{Der } B_m, \quad B_m = F[x_1, \dots, x_m]/(x_1^p, \dots, x_m^p).$$

$W(m : 1)$ may be graded and filtered by setting $\deg x_i = -\deg D_i = 1$ for $1 \leq i \leq m$. Then

$$W(m : 1) = \sum_{i \geq -1} W(m : 1)_{[i]}, \quad W(m : 1)_{[-1]} = \text{span}\{D_1, \dots, D_m\},$$

$$W(m : 1)_{[0]} = \text{span}\{x_i D_j | 1 \leq i, j \leq m\} \cong \mathfrak{gl}(W(m : 1)_{[-1]}).$$

$W(m : 1)$ is simple of dimension mp^m .

Albert and Frank [1] and Frank [12, 13] discovered restricted graded simple subalgebras M of $W(m : 1)$ obtained by taking $M = L^{(2)}$ where L is defined by $L_{[-1]} = W(m : 1)_{[-1]}$; $L_{[0]} \cong \mathfrak{sl}(m)$ (Frank [12]), $\mathfrak{sp}(m)$ for m even (Albert and Frank [1]), or $W(r : 1) + B_r$ acting on B_r for $m = p^r$ (Frank [13]) and

$$L_{[i]} = \{x \in W(m : 1)_{[i]} \mid (\mathrm{ad} L_{[-1]})^i x \subseteq L_{[0]}\} \quad \text{for } i \geq 1.$$

Simple nonrestricted subalgebras of $W(m : 1)$ were identified by Albert and Frank [1], Ree [30], Jennings and Ree [17], Strade [35].

All these families of nonclassical algebras are now known to be contained in the family of algebras of *Cartan type*. In 1966 Kostrikin and Šafarevič [26] constructed restricted simple Lie algebras corresponding to the infinite Lie algebras of Cartan over \mathbb{C} and conjectured that the algebras so constructed were precisely the known restricted simple Lie algebras over F . (Only the isomorphism of the algebras $K(m : 1)$ of Cartan type with one of the families discovered by Frank [13] was in question. This was proved in [10].) They also conjectured that all restricted simple Lie algebras over F were of classical or Cartan type.

The original Kostrikin-Šafarevič construction has been generalized to give nonrestricted algebras by Kostrikin-Šafarevič [27], Kac [19, 20, 21], Shen [33], Wilson [39, 40]. We describe the most general version [21, 40] here.

Give the polynomial algebra $F[X_1, \dots, X_m]$ its usual coalgebra structure with each X_i primitive. Then the dual space $\mathfrak{a}(m) = F[X_1, \dots, X_m]^*$ is an infinite-dimensional commutative associative algebra consisting of all formal sums $\sum x^\alpha$, where α ranges over all m -tuples of nonnegative integers, with multiplication determined by $x^\alpha x^\beta = \binom{\alpha+\beta}{\alpha} x^{\alpha+\beta}$ where $\binom{\gamma}{\alpha} = \binom{\gamma(1)}{\alpha(1)} \cdots \binom{\gamma(m)}{\alpha(m)}$. This algebra is called the completed free divided power algebra. For $\mathbf{n} = (n_1, \dots, n_m)$, an m -tuple of positive integers, we let $\mathfrak{a}(m : \mathbf{n})$ denote the span of the x^α with $\alpha(i) < p^{n_i}$ for all i . Then $\mathfrak{a}(m : \mathbf{n})$ is a subalgebra of $\mathfrak{a}(m)$. Write $\mathbf{1}$ for the m -tuple $(1, \dots, 1)$. Note that $\mathfrak{a}(m : \mathbf{1}) \cong B_m$. For each i we get a derivation D_i of $\mathfrak{a}(m)$ with $D_i(x^\alpha) = x^{\alpha - \epsilon_i}$, where $\epsilon_i(j) = \delta_{ij}$ (and $x^\beta = 0$ if some $\beta(i) < 0$). We will usually write x_i for x^{ϵ_i} . The set $\{u_1 D_1 + \cdots + u_m D_m \mid u_i \in \mathfrak{a}(m) \text{ (respectively, } u_i \in \mathfrak{a}(m : \mathbf{n}) \text{)}\}$ (where $(uD)v = u(Dv)$) is a subalgebra of $\mathrm{Der} \mathfrak{a}(m)$, which is denoted $W(m)$ (respectively $W(m : \mathbf{n})$). The algebra $W(m : \mathbf{n})$ is simple, of dimension mp^n , where $n = n_1 + \cdots + n_m$. It is restricted if and only if $\mathbf{n} = \mathbf{1}$.

Define differential forms $\omega_S, \omega_H, \omega_K$ by

$$\omega_S = dx_1 \wedge \cdots \wedge dx_m,$$

$$\omega_H = \sum_{i=1}^r dx_i \wedge dx_{i+r} \quad (m = 2r),$$

$$\omega_K = dx_{2r+1} + \sum_{i=1}^r x_{i+r} dx_i - x_i dx_{i+r} \quad (m = 2r + 1).$$

Elements of $W(m)$ act on differential forms according to the rules

$$D(df) = d(Df), \quad D(\alpha \wedge \beta) = (D\alpha) \wedge \beta + \alpha \wedge (D\beta),$$

$$D(f\alpha) = (Df)\alpha + f(D\alpha), \quad df = \sum_{i=1}^m (D_i f) dx_i$$

for all $f \in \mathfrak{a}(m)$ and all differential forms α and β . Define subalgebras $S(m)$, $H(m)$, $K(m) \subseteq W(m)$ by

$$S(m) = \{D \in W(m) \mid D(\omega_S) = 0\} \quad (m \geq 3),$$

$$H(m) = \{D \in W(m) \mid D(\omega_H) = 0\} \quad (m = 2r \geq 2),$$

$$K(m) = \{D \in W(m) \mid D(\omega_K) \in \mathfrak{a}(m)\omega_K\} \quad (m = 2r + 1 \geq 3).$$

For any automorphism Φ of $W(m)$ and for $X = S, H$ or K define

$$X(m : \mathbf{n} : \Phi) = \Phi X(m) \cap W(m : \mathbf{n}).$$

Then the algebras $W(m : \mathbf{n})$ and, for appropriate m , \mathbf{n} , and Φ depending on X (see [21]), the algebras $X(m : \mathbf{n} : \Phi)^{(2)}$ are simple. These are the algebras of Cartan type. Kac has shown [21] that the only restricted simple Lie algebras of Cartan type are those originally defined by Kostrikin and Šafarevič: $X(m : 1)^{(2)}$, $X = W, S, H, K$.

3. Classification results. We now state three recent classification results.

THEOREM 3.1 (WEISFEILER [38]). *Let L be finite-dimensional and simple over F and let L_0 be a maximal subalgebra of L which is solvable. Then $L \cong \mathfrak{sl}(2)$ or $W(1 : \mathbf{n})$ for some n .*

This generalizes earlier results of Schue [31] (where it is assumed that all proper subalgebras are solvable) and Kuznecov [28] (where it is assumed that L_0 acts irreducibly on L/L_0).

THEOREM 3.2 (BENKART-OSBORN [2, 3]). *Let L be finite-dimensional and simple over F and let L contain a one-dimensional Cartan subalgebra. Then $L \cong \mathfrak{sl}(2)$ or L is an Albert-Zassenhaus algebra (i.e., some $W(1 : \mathbf{n})$ or an appropriate $H(2 : \mathbf{n} : \Phi)^{(2)}$).*

This generalizes earlier work of Kaplansky [23] (where the theorem is proved for restricted L and structural results on nonrestricted algebras are obtained) and Block [5] (where the theorem is proved under the additional hypotheses that $\dim L_\alpha = 1$ and $[L_\alpha, L_{-\alpha}] \neq (0)$ for all roots α).

THEOREM 3.3 (BLOCK-WILSON [7, 8]). *Let L be a finite-dimensional restricted simple Lie algebra over F . Then L is of classical or Cartan type.*

This verifies the 1966 conjecture of Kostrikin and Šafarevič.

4. Techniques of proof. We now discuss briefly two techniques which are used in the proofs of the theorems cited above: filtration theory and local analysis.

Notice that the algebras of Cartan type all have natural filtrations by degree (in type K we take $\deg x_{2r+1} = 2$, in all other cases we take $\deg x_i = 1$). This, together with the heavy use of filtration theory in Cartan's work, suggests that we should try to equip an arbitrary simple algebra L with a filtration and use this filtration in our analysis.

Let L be simple and L_0 be a maximal subalgebra. Define a filtration (Cartan [9], Weisfeiler [36]) as follows: L_0 acts on L/L_0 . Let L_{-1}/L_0 be an irreducible L_0 -submodule. Set $L_{i+1} = \{x \in L_i \mid [x, L_{-1}] \subseteq L_i\}$ for $i \geq 0$ and $L_{i-1} = [L_i, L_{-1}] + L_i$ for $i < 0$. Let $G = \sum G_i$ be the associated graded algebra.

The aim of filtration theory is to determine first G and then L . The following "Recognition Theorem" which combines work of Kostrikin-Šafarevič [27], Kac [18, 21], and Wilson [40] shows that under appropriate hypotheses this can be done.

THEOREM 4.1. *Let L be finite-dimensional simple over F , filtered as above. Assume:*

(a) $G_0 \cong$ a direct sum of restricted ideals, each of which is either classical simple, $\mathfrak{gl}(k)$, $\mathfrak{sl}(k)$, or $\mathfrak{pgl}(k)$ where p divides k , or abelian.

(b) $x \in G_0$ implies $\text{ad}(x^p)$ and $(\text{ad } x)^p$ agree on G_{-1} .

(c) $x \in G_i$, $i \leq 0$, $[x, G_1] = (0)$ implies $x = 0$.

Then L is classical or of Cartan type (with the usual filtration).

A second major result in filtration theory (Weisfeiler [37]) gives the structure of G (without hypotheses on L_0). Before stating this result we recall Block's characterization of semisimple Lie algebras over F (which is used in Weisfeiler's result and also in local analysis).

THEOREM 4.2 (BLOCK [6]). *If L is a finite-dimensional semisimple Lie algebra over F then there exist integers $m, n_1, \dots, n_m \geq 0$ and simple Lie algebras S_1, \dots, S_m such that*

$$\sum_{i=1}^m S_i \otimes B_{n_i} \subseteq L \subseteq \text{Der} \left(\sum_{i=1}^m S_i \otimes B_{n_i} \right).$$

THEOREM 4.3 (WEISFEILER [37]). *Let L be finite-dimensional simple over F , filtered as above. Then $\text{solv } G \subseteq \sum_{i < 0} G_i$ and $G/\text{solv } G$ contains a unique minimal ideal $A \cong S \otimes B_n$, S simple, $n \geq 0$. Moreover, A is graded (with the grading determined by a grading on S if $A_1 \neq (0)$ and by a grading on B_n if $A_1 = (0)$) and $A_i = (G/\text{solv } G)_i$ for $i < 0$.*

In applications of this theorem one usually tries to show that $n = 0$ and that $\text{solv } G = (0)$. If this can be done it is frequently possible to apply Theorem 4.1.

Weisfeiler used these techniques to prove Theorem 3.1. The solvable maximal subalgebra L_0 is used to define the filtration. Then the solvable algebra G_0 acts irreducibly on G_{-1} . Representation theory (in particular, the theory of induced modules) allows one to describe G_{-1} well enough to show (after much work)

that if $G_{-2} \neq (0)$ then $\dim G = \infty$. Thus $L = L_{-1}$ and Kuznecov's result [28] applies.

The second major technique we will discuss, local analysis, involves studying subalgebras of L of toral rank 1 and 2 (the toral rank of algebra is the rank of the additive group generated by its roots) and using information obtained in this way to reconstruct L . We set

$$L^{(\alpha)} = \sum_{i \in \mathbb{Z}} L_{i\alpha}, \quad L[\alpha] = L^{(\alpha)} / \text{solv } L^{(\alpha)},$$

$$L^{(\alpha, \beta)} = \sum_{i, j \in \mathbb{Z}} L_{i\alpha + j\beta}, \quad L[\alpha, \beta] = L^{(\alpha, \beta)} / \text{solv } L^{(\alpha, \beta)}$$

for all roots α, β .

In the proof of Theorem 3.2, Benkart and Osborn observe that $L^{(\alpha)}$ contains a one-dimensional Cartan subalgebra Fh with all the eigenvalues of $\text{ad } h$ in the prime field. Algebras satisfying this condition have been classified by Yermolaev [43]. If such an algebra A is not solvable it has an abelian radical and $A / \text{solv } A \cong \text{sl}(2)$ or $W(1 : 1)$. This is enough information to develop a representation theory of these algebras. Benkart and Osborn do this and apply it to $L^{(\alpha)}$ acting on $\sum L_{\beta+i\alpha}$ for any root β . Information obtained in this way, together with use of filtration theory, allows them to show that the hypotheses of Block's theorem [5] are satisfied and so complete the proof.

In the proof of Theorem 3.3, the techniques of local analysis and filtration theory are combined. Local analysis is used to define an appropriate L_0 for use in Theorem 4.1.

Let A be any restricted Lie algebra and let T be a maximal torus in A . We say that T is *standard* in A if $C_A(T) = T + I$ where I is a nil ideal in $C_A(T)$. By [41] any maximal torus in a restricted simple Lie algebra L is standard.

Suppose L is restricted simple and T is a torus of maximal dimension in L . Then $M = L[\alpha]$ is restricted semisimple, has no tori of dimension > 1 (as T has maximal dimension), and all one-dimensional tori in M are standard (as all maximal tori in L are standard). An algebra M satisfying these conditions is of one of four types: $M = (0)$, $M \cong \text{sl}(2)$, $M \cong W(1 : 1)$, $H(2 : 1)^{(2)} \subseteq M \subseteq H(2 : 1)$. It is easily seen that in any case M contains a unique compositionally classical subalgebra of maximal dimension (see (2.6) for the case $W(1 : 1)$) and consequently $L^{(\alpha)}$ contains a unique compositionally classical subalgebra (which we denote $Q^{(\alpha)}$) of maximal dimension. If $T \subseteq Q^{(\alpha)}$ we say the root α is *proper*. We have seen (in (2.7)) that a root α can be either proper or improper. Use of Winter exponentials (see (2.8)) shows that if $\alpha \in T^*$ is improper we may switch to a new maximal torus T' such that the corresponding root $\alpha' \in T'^*$ is proper. (The rank one case can be completely analyzed for the nonrestricted case as well, but the list of possibilities is longer.)

Now consider the possible rank two algebras. We take T as above and further require that the number of proper roots with respect to T is maximal (so is at least 1). Then $M = L[\alpha, \beta]$ is restricted semisimple, has no tori of dimension

> 2 , and all two-dimensional tori in M are standard. Using Theorem 4.2 and the restrictedness of M (in particular, using Lemma 1.3) we see that either toral rank $M \leq 1$ or one of the following occurs:

(a) $S_1 \oplus S_2 \subseteq M \subseteq (\text{Der } S_1)^{(1)} \oplus (\text{Der } S_2)^{(1)}$ where $S_1, S_2 \cong \mathfrak{sl}(2)$, $W(1:1)$, or $H(2:1)^{(2)}$.

(b) $S \otimes B_n \subseteq M \subseteq \text{Der}(S \otimes B_n)$, $n > 0$, $S \cong \mathfrak{sl}(2)$, $W(1:1)$, or an appropriate $H(2:1:\Phi)^{(2)}$.

(c) $\bar{S} \subseteq M \subseteq \text{Der } S$ where \bar{S} denotes the restricted subalgebra of $\text{Der } S$ generated by S , S is simple, and $S + I \neq M$. Here $S \cong W(1:2)$ or an appropriate $H(2:1:\Phi)^{(2)}$.

(d) $\bar{S} \subseteq M \subseteq \text{Der } S$ where S is simple and $S + I = M$.

(The corresponding classification problem in the nonrestricted case is unsolved.)

To complete the analysis of case (d) we let $R(M)_\alpha = \{x \in M_\alpha \mid [x, M_{-\alpha}] \subseteq I\}$ and let $R(M) = C_M(T) + \sum R(M)_\alpha$. We let $M_0 \supseteq R(M)$ be a maximal subalgebra. Analysis of the rank one cases shows $\dim((M/M_0)_\gamma) \leq 7$ for all $\gamma \in T^*$ (this uses the hypothesis that rank $M = 2$), and the fact that there is at least one proper root with respect to T shows that $(M/M_0)_\gamma \neq (0)$ for at most $p^2 - p + 6$ values of γ . Then (filtering M as above and letting G be the associated graded algebra) the G_0 -module G_{-1} has the same properties. With this information we are able to show that Theorem 4.1 applies and hence that the simple algebra S in (d) is one of A_2 , C_2, G_2 , $W(2:1)$, $S(3:1)^{(1)}$, $H(4:1)^{(1)}$, $K(3:1)$.

Finally, consider a restricted simple algebra of arbitrary rank. Considering the subalgebra $(X(m:n:\Phi)^{(2)})_0$ of $X(m:n:\Phi)^{(2)}$, we observe that it contains a maximal torus T with respect to which all roots are proper and that, with respect to this torus, we have $(X(m:n:\Phi)^{(2)})_0 = \sum Q^{(\alpha)}$. This leads us to make the following claims:

(4.1) There is a torus T of maximal dimension in L with respect to which all roots are proper.

(4.2) $\sum Q^{(\alpha)}$ is a subalgebra (with $Q^{(\alpha)}$ taken with respect to T).

(4.3) If $L = \sum Q^{(\alpha)}$ then L is classical.

(4.4) If L_0 is a maximal subalgebra of L containing $\sum Q^{(\alpha)}$ and if $\Omega = \{\gamma \in \Gamma \mid L_\gamma \neq L_{0,\gamma}\}$ then for any $\alpha, \beta \in \Gamma$ we have $|(\beta + \mathbb{Z}\alpha) \cap \Omega| \leq 2$.

Using (4.4) we can show that L_0 satisfies the hypotheses of Theorem 4.1.

Properties (4.1)–(4.4) can all be checked in the rank two case. For example, for (4.1) it is sufficient to show that if α is improper and β proper, and if α', β' are corresponding roots under a Winter exponential with α' proper then β' is proper. If this property holds in every $L[\alpha, \beta]$ then it holds in L .

Unfortunately, these rank two properties do not hold in all of the semisimple algebras listed in (a)–(d). Thus we must show that some of these semisimple algebras cannot occur as $L[\alpha, \beta]$ for L simple. We give an example of how this can be done.

Suppose $L[\alpha, \beta] \cong \mathfrak{sl}(2) \otimes B_1 + F D_1 + F(x_1 D_1)$ and that $T = \text{span}\{h \otimes 1, x_1 D_1\}$ (where h is a nonzero diagonal matrix in $\mathfrak{sl}(2)$). Let α and $u \in L_\alpha$, $v \in L_{-\alpha}$

be such that $\bar{u} = u + \text{solv } L^{(\alpha, \beta)} = D_1$, $\bar{v} = h \otimes x_1$. Then $[\bar{u}, \bar{v}] = h \otimes 1$ and so $[u, v] \notin I$. However, $\alpha([u, L_{-\alpha}]) = (0)$. Let $\Gamma' = \{\gamma \in \Gamma \mid \gamma([u, v]) \neq 0\}$. Then $\sum_{\gamma \in \Gamma'} L_\gamma + \sum_{\gamma, \delta \in \Gamma'} [L_\gamma, L_\delta]$ is a nonzero ideal in L . Since L is simple this ideal equals L and so $C_L(T) = \sum_{\gamma \in \Gamma'} [L_\gamma, L_{-\gamma}]$. (This is an argument due to Schue [31].) Thus there exists $\gamma \in \Gamma$ satisfying $\gamma([u, v]) \neq 0$, $\alpha([L_\gamma, L_{-\gamma}]) \neq (0)$. From this it is easily seen that if J is any nonzero ideal in $M = L[\alpha, \gamma]$ then $M = J + I$. Furthermore, (letting $\bar{u} = u + \text{solv } L^{(\alpha, \gamma)}$ and $\bar{v} = v + \text{solv } L^{(\alpha, \gamma)}$) we have $\bar{u} \in M_\alpha$, $\bar{v} \in M_{-\alpha}$, $\gamma([\bar{u}, \bar{v}]) \neq 0$. However, $M[\alpha] = (0)$. Examination of (a)–(d) shows that there is no such M . Thus $L[\alpha, \beta]$ cannot have the given form.

Similar arguments allow us to eliminate all algebras for which (4.1)–(4.4) fail, and so Theorem 3.3 is proved.

REFERENCES

1. A. A. Albert and M. S. Frank, *Simple Lie algebras of characteristic p* , Univ. e Politec, Torino, Rend. Sem. Mat. **14** (1954–55), 117–139.
2. G. M. Benkart and J. M. Osborn, *On the determination of rank one Lie algebras of prime characteristic*, Contemp. Math., 13, Amer. Math. Soc., Providence, R.I., 1982, pp. 263–265.
3. —, *Rank one Lie algebras*, Ann. of Math. **119** (1984), 437–463.
4. R. E. Block, *New simple Lie algebras of prime characteristic*, Trans. Amer. Math. Soc. **89** (1958), 421–449.
5. —, *On Lie algebras of rank one*, Trans. Amer. Math. Soc. **112** (1964), 19–31.
6. —, *Determination of the differentiably simple rings with a minimal ideal*, Ann. of Math. **90** (1969), 433–459.
7. R. E. Block and R. L. Wilson, *The restricted simple Lie algebras are of classical or Cartan type*, Proc. Nat. Acad. Sci. U.S.A. **81** (1984), 5271–5274.
8. —, *Classification of the restricted simple Lie algebras*, Adv. in Math. (to appear).
9. E. Cartan, *Les groupes de transformations continus, infinis, simples*, Ann. Sci. École Normale (3) **26** (1909), 93–161; *Oeuvres complètes*, Vol. 2, Part II, Gauthier-Villars, Paris, 1953, pp. 857–925.
10. M. Ju. Celousov, *Derivations of Lie algebras of Cartan type*, Izv. Vyssh. Uchebn. Zaved. Mat. **98** (1970), 126–134. (Russian)
11. H. J. Chang, *Über Wittsche Lie-Ringe*, Abh. Math. Sem. Univ. Hamburg **14** (1941), 151–184.
12. M. S. Frank, *A new class of simple Lie algebras*, Proc. Nat. Acad. Sci. U.S.A. **40** (1954), 713–719.
13. —, *Two new classes of simple Lie algebras*, Trans. Amer. Math. Soc. **112** (1964), 456–482.
14. N. Jacobson, *Abstract derivation and Lie algebras*, Trans. Amer. Math. Soc. **42** (1937), 206–224.
15. —, *Classes of restricted Lie algebras of characteristic p* , II, Duke Math. J. **10** (1943), 107–121.
16. —, *Lie algebras*, Interscience, New York, 1962.
17. S. A. Jennings and R. Ree, *On a family of Lie algebras of characteristic p* , Trans. Amer. Math. Soc. **84** (1957), 192–207.
18. V. G. Kac, *The classification of the simple Lie algebras over a field with non-zero characteristic*, Izv. Akad. Nauk SSSR Ser. Mat. **34** (1970), 385–408; English transl. in Math. USSR-Izv. **4** (1970), 391–413.
19. —, *Global Cartan pseudogroups and simple Lie algebras of characteristic p* , Uspekhi Mat. Nauk **26** (1971), 199–200. (Russian)

20. —, *Filtered Lie algebras of Cartan type*, Uspekhi Mat. Nauk **29** (1974), 203–204. (Russian)
21. —, *Description of filtered Lie algebras with which graded Lie algebras of Cartan type are associated*, Izv. Akad. Nauk SSSR Ser. Mat. **38** (1974), 800–834; Errata **40** (1976), 1415; English transl. in Math. USSR-Izv. **8** (1974), 801–835; Errata **10** (1976), 1339.
22. I. Kaplansky, *Seminar on simple Lie algebras*, Bull. Amer. Math. Soc. **60** (1954), 470–471.
23. —, *Lie algebras of characteristic p* , Trans. Amer. Math. Soc. **89** (1958), 149–183.
24. S. Kobayashi and T. Nagano, *On filtered Lie algebras and geometric structures*. III, J. Math. Mech. **14** (1965), 679–706.
25. A. I. Kostrikin, *Variations modulaires sur un thème de Cartan*, Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 1, Gauthier-Villars, Paris, 1971, pp. 285–292.
26. A. I. Kostrikin and I. R. Šafarevič, *Cartan pseudogroups and Lie p -algebras*, Dokl. Akad. Nauk SSSR **168** (1966), 740–742; English transl. in Soviet Math. Dokl. **7** (1966), 715–718.
27. —, *Graded Lie algebras of finite characteristic*, Izv. Akad. Nauk SSSR Ser. Mat. **33** (1969), 251–322; English transl. in Math. USSR-Izv. **3** (1969), 237–304.
28. M. I. Kuznecov, *The modular simple Lie algebras with a solvable maximal subalgebra*, Mat. Sb. **101** (1976), 77–86; English transl. in Math. USSR-Sb. **30** (1976), 68–76.
29. W. H. Mills and G. B. Seligman, *Lie algebras of classical type*, J. Math. Mech. **6** (1957), 519–548.
30. R. Ree, *On generalized Witt algebras*, Trans. Amer. Math. Soc. **83** (1956), 510–546.
31. J. R. Schue, *Cartan decompositions for Lie algebras of prime characteristic*, J. Algebra **11** (1969), 25–52; Erratum **13** (1969), 558.
32. G. B. Seligman, *Modular Lie algebras*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 40, Springer-Verlag, New York, 1967.
33. G. Y. Shen, *New simple Lie algebras of characteristic p* , Chinese Ann. Math. Ser. B **4** (1983), 329–346.
34. I. M. Singer and S. Sternberg, *The infinite groups of Lie and Cartan*. I, *The transitive groups*, J. Analyse Math. **15** (1965), 1–114.
35. H. Strade, *Nodale nichtkommutative Jordanalgebren und Lie-Algebren bei Charakteristik $p > 2$* , J. Algebra **21** (1972), 353–377.
36. B. Ju. Weisfeiler, *Infinite dimensional filtered Lie algebras and their connection with graded Lie algebras*, Funktsional. Anal. i Prilozhen **2** (1968), 94–95. (Russian)
37. —, *On the structure of the minimal ideal of some graded Lie algebras in characteristic $p > 0$* , J. Algebra **53** (1978), 344–361.
38. —, *On subalgebras of simple Lie algebras of characteristic $p > 0$* , Trans. Amer. Math. Soc. **286** (1984), 471–503.
39. R. L. Wilson, *Nonclassical simple Lie algebras*, Bull. Amer. Math. Soc. **75** (1969), 987–991.
40. —, *A structural characterization of the simple Lie algebras of generalized Cartan type over fields of prime characteristic*, J. Algebra **40** (1976), 418–465.
41. —, *Cartan subalgebras of simple Lie algebras*, Trans. Amer. Math. Soc. **234** (1977), 435–446.
42. D. J. Winter, *On the toral structure of Lie p -algebras*, Acta Math. **123** (1969), 69–81.
43. J. B. Yermolaev, *Lie algebras of rank 1 with root system in the prime field*, Izv. Vyssh. Uchebn. Zaved. Mat. **120** (1972), 38–50. (Russian)
44. H. Zassenhaus, *Über Liesche Ringe mit Primzahlcharakteristik*, Abh. Math. Sem. Univ. Hamburg **13** (1939), 1–100.

Kloosterman Zeta Functions for $GL(n, \mathbf{Z})$

DORIAN GOLDFELD

1. Introduction. Analytic number theory has made considerable strides in the past few years. Let me begin by citing two of the most striking recent results.

Fouvry [9] has shown that there exist infinitely many primes p such that $p-1$ has a prime factor larger than $p^{2/3}$. This, in conjunction with a theorem of L. M. Adleman and D. R. Heath-Brown [1] (extensions of Sophie Germaine's criterion), enables one to show that Fermat's equation

$$x^p + y^p = z^p \quad (p \nmid xyz)$$

has no positive integral solutions for infinitely many primes p .

Another major breakthrough has been the work of Deshouillers-Iwaniec [8] which has led to many spectacular results. For example, Bombieri-Friedlander-Iwaniec [2, 3] have recently obtained an averaged form of the prime number theorem for arithmetic progressions (of Bombieri-Vinogradov type). In the case of the distribution of primes $\leq x$ with respect to moduli $> \sqrt{x}$, their results go beyond what can be obtained upon assumption of the generalized Riemann hypothesis.

The proofs of the above theorems have a new common ingredient; uniform estimates (of the type first proved by Kuznetsov [16]) for the distribution of Kloosterman sums. For other applications of this innovative idea, the excellent survey article of Iwaniec [13] is to be commended. Accordingly, our attention is turned to a general theory of Kloosterman sums, hyper-Kloosterman sums, and their zeta functions.

Let $M, N \in \mathbf{Z}$ and $s \in \mathbf{C}$. The Kloosterman zeta function for $GL(2, \mathbf{Z})$ is defined to be

$$Z(M, N; s) = \sum_{c=1}^{\infty} S(M, N; c) c^{-2s} \quad (1.1)$$

where

$$S(M, N; c) = \sum_{\substack{a=1 \\ (a,c)=1}}^c e^{2\pi i(aM + \bar{a}N)/c} \quad (a\bar{a} \equiv 1 \pmod{c}) \quad (1.2)$$

Supported in part by National Science Foundation Grant DMS-8641838.

is the classical Kloosterman sum. The bound $S(M, N; c) = O(c^{1/2+\varepsilon})$ of A. Weil [22] shows that (1.1) converges absolutely for $\operatorname{Re}(s) > \frac{3}{4}$.

The function (1.1) was first introduced by A. Selberg [19] who obtained its meromorphic continuation to the whole complex s -plane. In [12], by use of bounds for the resolvent operator, Goldfeld and Sarnak have shown that

$$Z(M, N; s) = O(|s|^{1/2+\varepsilon}) \quad (1.3)$$

for $\operatorname{Re}(s) > \frac{1}{2} + \varepsilon$ and $|\operatorname{Im}(s)| > \varepsilon$. The bound (1.3) leads to a simple proof of Kuznetsov's theorem [16]

$$\sum_{c \leq x} \frac{S(M, N; c)}{c} = O(x^{1/6+\varepsilon}) \quad (1.4)$$

which we discussed before. Other generalizations of (1.4) have been obtained by Deshouillers and Iwaniec [8] and Proskurin [18]. Also, Bruggeman [4] has developed a Kuznetsov trace formula which also leads to (1.4).

We shall now consider Kloosterman zeta functions for higher rank groups, focussing on $\operatorname{GL}(n, \mathbf{Z})$ with $n > 2$. Uniform estimates for the distribution of hyper-Kloosterman sums and products of classical Kloosterman sums should be the outcome of this endeavor.

2. Notation. For $n = 2, 3, \dots$ let $G = \operatorname{GL}(n, \mathbf{R})$, $\Gamma = \operatorname{GL}(n, \mathbf{Z})$, $X \subset G$ be the set of all upper triangular matrices with ones on the diagonal, and $Y \subset G$ the set of diagonal matrices of type

$$\operatorname{diag}(y_1 \cdots y_{n-1}, y_1 \cdots y_{n-2}, \dots, y_1, 1) \quad (2.1)$$

with $y_i > 0$. We consider the homogeneous space $H \cong G/O(n, \mathbf{R}) \cdot \mathbf{R}$ where $O(n, \mathbf{R})$ is the orthogonal group. By the Iwasawa decomposition, every $z \in H$ has a unique decomposition $z \equiv xy \pmod{O(n, \mathbf{R}) \cdot \mathbf{R}}$ with $x \in X$ and $y \in Y$. The discrete group Γ acts on H by left matrix multiplication. Let $\mathcal{L}^2(\Gamma \backslash H)$ denote the Hilbert space with inner product

$$\langle f, g \rangle = \int_{\Gamma \backslash H} f(z) \overline{g(z)} d^*z$$

where both $f, g: H \rightarrow \mathbf{C}$ are left-invariant under Γ and the invariant volume element d^*z satisfies

$$d^*z = \prod_{1 < i < j \leq n} dx_{i,j} \prod_{i=1}^{n-1} y_i^{-i(n-i)-1} dy_i$$

where $x = (x_{ij}) \in X$, $y \in Y$ is given by (2.1) and $z \equiv xy$.

Henceforth $M = (M_1, \dots, M_{n-1})$, $N = (N_1, \dots, N_{n-1})$ are in \mathbf{Z}^{n-1} and $s = (s_1, \dots, s_{n-1})$, $u = (u_1, \dots, u_{n-1})$, $v = (v_1, \dots, v_{n-1})$, and $k = (k_1, \dots, k_{n-1})$ are in \mathbf{C}^{n-1} . By θ_M , θ_N , we mean characters of X given by

$$\theta_M(x) = e(M_1 x_{1,2} + \cdots + M_{n-1} x_{n-1,n})$$

with $x = (x_{ij}) \in X$ and $e(\theta) = e^{2\pi i \theta}$.

3. Kloosterman sums associated to double coset decompositions of $GL(n, \mathbf{Z})$. Let W denote the Weyl group of G . If $D \subset G$ is the subgroup of diagonal matrices, then we have the Bruhat decomposition $G = \bigcup_{w \in W} XDwX$. This induces the decomposition $\Gamma = \bigcup_{w \in W} \Gamma_w$ where $\Gamma_w = (XDwX) \cap \Gamma$ are termed Bruhat cells. The cell corresponding to the so called long element

$$w = \begin{pmatrix} & & & \pm 1 \\ & & 1 & \\ & \ddots & & \\ 1 & & & \end{pmatrix} \quad (3.1)$$

is called the big cell.

Consider the minimal parabolic subgroup $P = X \cap \Gamma$ of Γ , and the various subgroups $P_w = ((w^{-1})^t P w) \cap P$ indexed by $w \in W$. Let (c_1, \dots, c_{n-1}) be nonzero integers and set

$$c = \text{diag}(1/c_{n-1}, c_{n-1}/c_{n-2}, \dots, c_2/c_1, c_1). \quad (3.2)$$

For $M, N \in \mathbf{Z}^{n-1}$, the generalized Kloosterman sum $S_w(M, N; c)$ is defined as

$$S_w(M, N; c) = \sum_{\substack{\gamma \in P \backslash \Gamma_w / P_w \\ \gamma = b_1 c w b_2}} \theta_M(b_1) \theta_N(b_2) \quad (3.3)$$

where θ_M, θ_N are characters of X . This reduces to the classical sum (1.2) for $n = 2$ and $w = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}$.

The sum (3.3) was first considered in Bump-Friedberg-Goldfeld [6, 7] for $n = 3$, and somewhat later for $n > 3$ by Friedberg [10], Stevens [21] and Piatetski-Shapiro. As shown in [10], the Kloosterman sums are multiplicative in c , nonzero only if $w \in W$ is of the form

$$(\text{mod } SL(n, \mathbf{Z}) \cap D), \quad w = \begin{pmatrix} & & & I_1 \\ & & I_2 & \\ & \ddots & & \\ I_l & & & \end{pmatrix}$$

where the I_j are identity matrices; and factor into nondegenerate classical Kloosterman sums of type (1.2) if c_1, \dots, c_{n-1} are pairwise coprime and w is the long element (3.1). If c is given by (3.2) and c_i are suitable powers of a fixed prime p , then (3.3) will be associated to an algebraic variety over \mathbf{F}_p . In the special case

$$c = (p^{1-n}, p, \dots, p), \quad w = \begin{pmatrix} & & \pm 1 \\ & & \\ & & \\ I_{n-1} & & \end{pmatrix},$$

[10] has shown that $S_w(M, N; c)$ is a power of p times

$$\sum_{x_1 \cdots x_{n-1} \equiv M_1 \cdots M_{n-1} N_{n-1} \pmod{p}} e((x_1 + \cdots + x_{n-1})/p)$$

and is, therefore, associated to a Kloosterman hypersurface. In general (see [21]), the associated varieties are not smooth and their classification is still an open problem.

4. Kloosterman zeta functions. Let $M, N \in \mathbf{Z}^{n-1}$,

$$c = \text{diag}(1/c_{n-1}, c_{n-1}/c_{n-2}, \dots, c_2/c_1, c_1),$$

$w \in W$, and $s \in \mathbf{C}^{n-1}$. The Kloosterman zeta function associated to the Bruhat cell Γ_w is defined to be

$$Z_w(M, N; s) = \sum_{c_1=1}^{\infty} \cdots \sum_{c_{n-1}=1}^{\infty} S_w(M, N; c) c_1^{-ns_1} \cdots c_{n-1}^{-ns_{n-1}}. \quad (4.1)$$

For $n > 2$, little is known about this function at present. We expect, however, that (4.1) has a meromorphic continuation in s , and that the polar divisors of (4.1) may be a subset of the polar divisors of the global zeta function

$$Z(M, N; s) = \sum_{w \in W} Z_w(M, N; s). \quad (4.2)$$

If this were the case for $n = 3$, then by the arguments of [6, 7] the generalized Ramanujan conjecture would follow for $n = 3$, and by the Gelbart-Jacquet lift [11], also for $n = 2$.

The meromorphic continuation of (4.2) has been obtained for $n = 3$ in [6, 7] by considering the inner product of two Poincaré series. We indicate an approach to generalizing our results to $n > 3$.

For $v \in \mathbf{C}^{n-1}$, define $I_v: H \rightarrow \mathbf{C}$ by the formula

$$I_v(z) = \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} y_i^{b_{ij}v_j}, \quad (4.3)$$

$$b_{ij} = \begin{cases} (n-i)j, & 1 \leq j \leq i, \\ (n-j)i, & i \leq j \leq n-i, \end{cases}$$

where $z \equiv xy$ with $x \in X$ and $y \in Y$ given by (2.1). Let \mathcal{D} denote the polynomial ring of differential operators defined on $\Gamma \backslash H$. Every $d \in \mathcal{D}$ determines a character $\lambda_v(d)$ given by $dI_v = \lambda_v(d)I_v$.

A Maass form of type $k \in \mathbf{C}^{n-1}$ is a smooth function $\varphi \in \mathcal{L}^2(\Gamma \backslash H)$ satisfying $d\varphi = \lambda_k(d)\varphi$ for all $d \in \mathcal{D}$. If it is "cuspidal" it has a Whittaker expansion [17, 20]

$$\varphi(z) = \sum_{M_1=1}^{\infty} \cdots \sum_{M_{n-1}=1}^{\infty} \sum_{\gamma \in \Lambda_{n-1}} a_M \prod_{i=1}^{n-1} M_i^{(-i(n-i))/2} W_k \left((M) \begin{pmatrix} \gamma & 0 \\ 0 & 1 \end{pmatrix} z \right) \quad (4.4)$$

where $M = (M_1, \dots, M_{n-1})$,

$$(M) = \text{diag}(M_1 \cdots M_{n-1}, M_1 \cdots M_{n-2}, \dots, M_1, 1),$$

$a_M \in \mathbf{C}$, $\Lambda_{n-1} = U_{n-1} \backslash GL(n-1, \mathbf{Z})$, $U_{n-1} \subset SL(n-1, \mathbf{Z})$ is the subgroup of upper triangular matrices with ones on the diagonal, and $W_k(z)$ is the Whittaker function given by

$$W_k(z) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_k(w_o u z) e \left(- \sum_{j=1}^{n-1} u_{j,j+1} \right) \cdot \prod_{1 \leq i < j \leq n} du_{ij}$$

where w_o is the long element (3.1) and $u = (u_{ij}) \in X$. Although we have not defined "cuspidal," we may take (4.4) as a definition.

The meromorphic continuation in $s \in \mathbf{C}^{n-1}$ of the Mellin transform of the Whittaker function

$$\mathcal{M}_k(s) = \int_0^\infty \cdots \int_0^\infty y_1^{s_1} \cdots y_{n-1}^{s_{n-1}} W_k(y) \prod_{i=1}^{n-1} \frac{dy_i}{y_i}$$

is still unknown except for $n = 2, 3$ (see Bump [5]). A heuristic argument of Ka-Lam Kueh suggests that $\mathcal{M}_k(s)$ has its first simple poles at

$$s_i = \left(\sum_{j=1}^{n-1} b_{n-i,j} k_j \right) - i(n-i) \quad (i = 1, \dots, n-1) \quad (4.5)$$

with b_{ij} given by (4.3).

Let $N \in \mathbf{Z}^{n-1}$ and θ_N be a character of X . An e -function is a bounded function $e_N: H \rightarrow \mathbf{C}$ satisfying $e_N(uz) = \theta_N(u)e_N(z)$ for all $u \in X$ and $z \in H$. Let $v \in \mathbf{C}^{n-1}$. We consider the Poincaré series

$$P_N(z; v) = \sum_{\gamma \in P \backslash \Gamma} I_v(\gamma z) e_N(\gamma z).$$

For $u, v \in \mathbf{C}^{n-1}$, $M, N \in \mathbf{Z}^{n-1}$, the inner product of $P_M(z, v)$ and $P_N(z, u)$ satisfies

$$\begin{aligned} \langle P_M, P_N \rangle &= \sum_{w \in W} \sum_c \frac{S_w(M, N; c)}{c^{nv}} \\ &\quad \cdot \int_{X_w} \int_{y_1=0}^{\infty} \cdots \int_{y_{n-1}=0}^{\infty} I_v(wz) e_M(cwz) \overline{I_u(z)} \overline{e_N(z)} d^* z \end{aligned} \quad (4.6)$$

where $X_w = ((w^{-1})^t X w) \cap X$, $c^{nv} = c_1^{nv_1} \cdots c_{n-1}^{nv_{n-1}}$, and the sum on the right side of (4.6) goes over all $N = (\varepsilon_1 N_1, \dots, \varepsilon_{n-1} N_{n-1})$ with $\varepsilon_i = \pm 1$ and $\varepsilon_1 \cdots \varepsilon_{n-1} = 1$.

For fixed u , the integrals on the dexter side of (4.6) can be continued as analytic functions of v . The polar set of $Z(M, N; v)$ can then be obtained from the polar set of $\langle P_M, P_N \rangle$. Let φ be a Maass form of type $k \in \mathbf{C}^{n-1}$. The projection of P_M, P_N onto φ yields a contribution $\langle P_M, \varphi \rangle \cdot \langle \overline{P_N}, \overline{\varphi} \rangle$ to $\langle P_M, P_N \rangle$. Some formal calculations in conjunction with (4.4) give

$$\langle P_M, \varphi \rangle = \bar{a}_M \left[\prod_{i=1}^{n-1} M_i^{(i(n-i)/2 - \sum_{j=1}^{n-1} b_{n-i,j} v_j)} \right] \mathcal{M}_{\bar{k}}(t)$$

with $t = (t_1, \dots, t_{n-1})$ and $t_i = \left(\sum_{j=1}^{n-1} b_{n-i,j} v_j \right) - i(n-i)$. Since $P_M(z, v)$ is orthogonal to the residual spectrum of \mathcal{D} , we do not obtain residual polar divisors. By (4.5), we are led to expect that $Z(M, N; s)$ has a meromorphic continuation in s with simple polar divisors which contain the hyperplanes

$$\sum_{j=1}^{n-1} b_{n-i,j} s_j = \sum_{j=1}^{n-1} b_{ij} \bar{k}_j \quad (i = 1, \dots, n-1).$$

5. Kloosterman decompositions of Selberg's kernel function. Another approach to the distribution of Kloosterman sums is based on a double coset decomposition of Selberg's kernel function. This method was used by Zagier (see Iwaniec [14]) to give an alternate proof of Kuznetsov's sum formula, and more recently by Ye [23] to give a new proof of quadratic base change. We consider generalizations to $G = \mathrm{GL}(n, \mathbf{R})$ with $n \geq 2$ which lead to new types of trace formulae.

Consider the Cartan decomposition $G = KAK$ where $K = \mathrm{O}(n, \mathbf{R})$ and $A \subset G$ is the subgroup of diagonal matrices with positive entries. Let $\varphi: K \backslash G / K \rightarrow \mathbf{C}$ be a K -biinvariant function.

Formally, the Selberg kernel function for Γ is

$$K(z, z') = \sum_{\gamma \in \Gamma} \varphi(z^{-1} \gamma z') \quad (5.1)$$

where $z, z' \in H$. For suitably chosen φ the dexter side of (5.1) converges absolutely and uniformly on compact subsets of $H \times H$.

Now (5.1) can be rewritten

$$K(z, z') = \sum_{(m) \in P} \sum_{\gamma \in P \backslash \Gamma} \varphi((m)z)^{-1} \gamma z') \quad (5.2)$$

with $(m) = (m_{ij})$. For fixed $(m) \in P$, the inner sum on the dexter side of (5.2), denoted $K_{(m)}(z, z')$ is an automorphic form for $\Gamma \backslash H$. It has a Fourier expansion in x with N th Fourier coefficient (here $N \in \mathbf{Z}^{n-1}$) given by

$$\int_{P \backslash X} K_{(m)}(z, z') \overline{\theta_N(x)} dx \quad (5.3)$$

which is itself a Poincaré series in z' . For $M \in \mathbf{Z}^{n-1}$, the M th Fourier coefficient in x' of (5.3) is

$$\int_{P \backslash X} \int_X \sum_{\gamma \in P \backslash \Gamma} \varphi(z^{-1} \gamma z') \overline{\theta_N(x)} \overline{\theta_M(x')} dx dx',$$

which is just

$$\sum_{w \in W} \sum_c S_w(M, N; c) \int_{X_w} \int_X \varphi(z^{-1} c w z') \overline{\theta_N(x)} \overline{\theta_M(x')} dx dx' \quad (5.4)$$

with the notation of (4.6).

For $x \in \mathbf{C}$, $z \in H$, and P_0 the maximal parabolic subgroup $(0 \dots 01)^*$ of $SL(n, \mathbf{Z})$, let $E(z, s) = \sum_{\gamma \in P_0 \backslash \Gamma} (\det \gamma z)^s$ denote the maximal parabolic Eisenstein series which converges absolutely and uniformly on compact subsets of H for $\operatorname{Re}(s) > 1$. Now, if $K_0(z, z')$ is the projection of $K(z, z')$ onto the space of cuspidal Maass forms, we are interested in computing the trace

$$\operatorname{Res}_{s=1} \int_{\Gamma \backslash H} K_0(z, z) E(z, s) d^* z. \quad (5.5)$$

After some formal computations (see Jacquet [15]), it follows from (5.4) that the essential contribution to the trace (5.5) is given by

$$\operatorname{Res}_{s=1} \sum_{w \in W} \sum_c \sum_{N \neq (0)} S_w(N, N; c) F_w(N, c, s) \quad (5.6)$$

where

$$F_w = \int_{y_1=0}^{\infty} \cdots \int_{y_{n-1}=0}^{\infty} \int_{X_w} \int_X \varphi(y^{-1} x^{-1} c w x' y) \theta_N(x + x') dx dx' \\ \cdot \prod_{i=1}^{n-1} y_i^{(s-1)(n-i)-1} dy_i.$$

In the special case $n = 2$, (5.6) takes the form

$$\operatorname{Res}_{s=1} \sum_{N \neq 0} \sum_{c=1}^{\infty} \frac{S(N, N; c)}{c^{s+1}} F\left(\frac{N}{c}, s\right)$$

where

$$F(B, s) = \int_0^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi \left(\begin{pmatrix} 1 & \\ x & 1 \end{pmatrix} \begin{pmatrix} y & \\ & y^{-1} \end{pmatrix} \begin{pmatrix} 1 & x' \\ & 1 \end{pmatrix} \right) \\ \cdot e(-By(x + x')) y^s dx dx' dy$$

decays rapidly to zero as $B \rightarrow \infty$.

To compute the above residue we use a method of Kuznetsov. Let

$$S(N, N; c) = \sum_{-c/2 \leq l \leq c/2} v(c, l) e\left(\frac{lN}{c}\right)$$

where $v(c, l)$ denotes the number of solutions $a \pmod{c}$ of $a^2 - al + 1 \equiv 0 \pmod{c}$. By Poisson summation

$$\sum_N e\left(\frac{lN}{c}\right) F\left(\frac{N}{c}, s\right) = \sum_{h \in \mathbf{Z}} \int_{-\infty}^{\infty} F\left(\frac{\xi}{c}, s\right) e\left(\left(\frac{l}{c} + h\right)\xi\right) d\xi$$

where the integral on the right is bounded by $(hc)^{-m}$ for $h \neq 0$ and bounded by l^{-m} for $h = 0$, $l \neq 0$ (after integrating by parts $m > 2$ times). Consequently, the residue (5.7) is given by

$$\operatorname{Res}_{s=1} \left[\sum_{c=1}^{\infty} \sum_{l=-\infty}^{\infty} \frac{v(c, l)}{c^s} \right] \cdot \int_{-\infty}^{\infty} F(\xi, s) e(-l\xi) d\xi.$$

We have, therefore, expressed the principal cuspidal contribution to the Selberg trace formula in terms of special values of quadratic L -functions.

REFERENCES

1. L. M. Adleman and D. R. Heath-Brown, *The first case of Fermat's last theorem*, Invent. Math. **79** (1985), 409–416.
2. E. Bombieri, J. Friedlander, and H. Iwaniec, *Primes in arithmetic progression for large moduli*. I, Acta Math. **156** (1986), nos. 3–4, 203–251.
3. —, *Primes in arithmetic progression for large moduli*. II, Math. Ann. (to appear).
4. R. W. Bruggeman, *Fourier coefficients of automorphic forms*, Lecture Notes in Math., vol. 865, Springer-Verlag, Berlin and New York, 1981.
5. D. Bump, *Automorphic forms on $GL(3, \mathbb{R})$* , Lecture Notes in Math., vol. 1083, Springer-Verlag, Berlin and New York, 1984.
6. D. Bump, S. Friedberg, and D. Goldfeld, *Poincaré series and Kloosterman sums*, Contemporary Math., vol. 53, Amer. Math. Soc., Providence, R.I., 1986, pp. 39–49.
7. —, *Poincaré series and Kloosterman sums for $GL(3, \mathbb{Z})$* , Acta Arith. (to appear).
8. J. M. Deshouillers and H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms*, Invent. Math. **70** (1982), 219–288.
9. E. Fouvry, *Théorème de Brun-Titchmarsh; application au théorème de Fermat*, Invent. Math. **79** (1985), 383–408.
10. S. Friedberg, *Poincaré series for $GL(n)$: Fourier expansion, Kloosterman sums, and algebro-geometric estimates* (to appear).
11. S. Gelbart and H. Jacquet, *A relation between automorphic representations of $GL(2)$ and $GL(3)$* , Ann. Sci. École Norm. Sup. **11** (1978), 471–542.
12. D. Goldfeld and P. Sarnak, *Sums of Kloosterman sums*, Invent. Math. **71** (1983), 243–250.
13. H. Iwaniec, *Non-holomorphic modular forms and their applications*, Modular Forms (R. Rankin, ed.), Ellis Horwood, West Sussex, 1984, pp. 157–197.
14. —, *Promenade along modular forms and number theory*, Topics in Analytic Number Theory, Univ. Texas Press, Austin, Texas, 1985, pp. 221–303.
15. H. Jacquet, *Dirichlet series for the group $GL(n)$* , Automorphic Forms, Representation Theory and Arithmetic, Bombay Colloquium, Springer-Verlag, 1981, pp. 155–164.
16. N. V. Kuznetsov, *Petersson hypothesis for parabolic forms of weight zero and linnik hypothesis*, Sums of Kloosterman Sums, Mat. Sb. **111(153)** (1980), 334–383.
17. I. I. Piatetski-Shapiro, *Euler subgroups*, Lie Groups and Their Representations, Wiley, New York, 1975, pp. 597–620.
18. N. V. Proskurin, *Summation formulas for generalized Kloosterman sums*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. **82** (1979), 103–135.
19. A. Selberg, *On the estimation of Fourier coefficients of modular forms*, Proc. Sympos. Pure Math., no. 8, Amer. Math. Soc., Providence, R.I., 1965, pp. 1–15.
20. J. Shalika, *The multiplicity one theorem for $GL(n)$* , Ann. of Math. **100** (1974), 171–193.
21. G. Stevens, *On the Fourier coefficients of Poincaré series on $GL(r)$* (to appear).
22. A. Weil, *On some exponential sums*, Proc. Nat. Acad. Sci. U.S.A. **34** (1948), 204–207.
23. Y. Ye, *Kuznetsov trace formula and base change*, Ph. D. Thesis, Columbia University, New York, 1986.

COLUMBIA UNIVERSITY, NEW YORK, NEW YORK 10027, USA

Heights and L -Series

BENEDICT H. GROSS

One prominent aspect of current research in number theory has been the attempt to relate the L -series of certain automorphic forms to the L -series of motives. Once established, such relations usually lead to a fruitful interplay between the analytic and arithmetic sides of the subject, in the spirit of Dirichlet's class number formula and Kronecker's limit formula.

Over the last thirty years great advances have been made in establishing a correspondence between the L -series of holomorphic modular forms for $GL(2)$ and the L -series of 2-dimensional l -adic Galois representations. Zagier and I have recently found a limit formula in this setting, which I will discuss in a simple case below. This formula connects the heights of special divisor classes on modular curves to the behavior of certain Rankin L -series in the center of their critical strip. This paper is expository in nature, and is intended as an introduction to the main results and the steps in the proofs. For the details, we refer the reader to the papers in our bibliography.

1. Dirichlet L -series. Let K be an imaginary quadratic field of discriminant $-p$, where p is a prime with $p \equiv 3 \pmod{4}$ and $p > 3$. Let \mathcal{O} denote the ring of integers in K , and for $n \geq 1$ let $R(n)$ be the number of ideals of index n in \mathcal{O} . The Riemann zeta function is defined by the series $\zeta(s) = \sum_{n \geq 1} n^{-s}$ and the zeta function of K is defined by the analogous series $\zeta_K(s) = \sum_{n \geq 1} R(n)n^{-s}$. Both Dirichlet series converge absolutely, and define analytic functions of s , in the half plane $\operatorname{Re}(s) > 1$.

We define a periodic function ε on \mathbf{Z} by $\varepsilon(n) = \left(\frac{n}{p}\right)$ for $(n, p) = 1$ and $\varepsilon(n) = 0$ if $n \equiv 0 \pmod{p}$. If l is a prime, we have $\varepsilon(l) = 1$ if and only if l splits in K , by the theorem of quadratic reciprocity. Define the Dirichlet series $L(\varepsilon, s) = \sum_{n \geq 1} \varepsilon(n)n^{-s}$. Using the Euler products of the three series we have defined, one can obtain Dirichlet's factorization

$$\zeta_K(s) = \zeta(s)L(\varepsilon, s). \quad (1.1)$$

Comparing the coefficients of n^{-s} on both sides of (1.1) gives the formula

$$R(n) = \sum_{d|n} \varepsilon(d). \quad (1.2)$$

A much deeper corollary of (1.1) is Dirichlet's formula for the class number h of K . This results from a comparison of the residue of $\zeta_K(s)$ at $s = 1$ with the value $L(\varepsilon, 1)$. In the case we are considering, the class number formula can be written

$$h = -\frac{1}{p} \sum_{a=1}^p \left(\frac{a}{p} \right) a. \quad (1.3)$$

2. Rankin L -series. We now turn to the L -series of modular forms. Let N be a prime with $N \neq p$ and let $f(z) = \sum_{n \geq 1} a_n e^{2\pi i n z}$ be a holomorphic cusp form of weight 2 for the group $\Gamma_0(N)$. Define the twisted form $f \otimes \varepsilon$ by the sum $\sum_{n \geq 1} a_n \varepsilon(n) e^{2\pi i n z}$; this is a cusp form of weight 2 for the group $\Gamma_0(Np^2)$. The Rankin L -function we will study is defined by the product

$$L_K(f, s) = \sum_{\substack{n \geq 1 \\ (n, N)=1}} \varepsilon(n) n^{1-2s} \cdot \sum_{n \geq 1} a_n R(n) n^{-s}, \quad (2.1)$$

which converges in the half-plane $\operatorname{Re}(s) > \frac{3}{2}$.

We say f is a newform if it is an eigenform for the Hecke operators and is normalized by the condition $a_1 = 1$. (Since there are no cusp forms of weight 2 for $\operatorname{SL}_2(\mathbf{Z})$, there are no old eigenforms of prime level N .) In the case when f is a newform, the Rankin L -series defined by (2.1) has an Euler product (of degree 4) and factors as in (1.1):

$$L_K(f, s) = \sum_{n \geq 1} a_n n^{-s} \sum_{n \geq 1} a_n \varepsilon(n) n^{-s} = L(f, s) L(f \otimes \varepsilon, s). \quad (2.2)$$

The newform f corresponds to a new vector in an automorphic representation π_f of GL_2 over \mathbf{Q} , and $L_K(f, s)$ is the (finite part of the) L -function of the lifting of π_f to K .

When f is a newform, we can also give a motivic interpretation of the function $L_K(f, s)$. Indeed, there is a 2-dimensional l -adic representation

$$\sigma_f: \operatorname{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \operatorname{GL}_2(\overline{\mathbf{Q}}_l)$$

which is unramified at all primes $q \neq l, N$. Moreover, the image of a Frobenius element in the decomposition group of q satisfies the characteristic polynomial $x^2 - a_q x + q = 0$. The series $L_K(f, s)$ is the (finite part of the) L -function of the 4-dimensional l -adic representation $\sigma_f \otimes \operatorname{Ind}_{\mathbf{Q}}^{\mathbf{Q}_l} 1$. It can also be viewed as the L -series of the 2-dimensional representation σ_f , restricted to the Galois group $\operatorname{Gal}(\overline{\mathbf{Q}}/K)$. If f has rational Fourier coefficients, it corresponds to an elliptic curve E up to isogeny over \mathbf{Q} . The representation σ_f arises from the Galois action on l -power division points of E , and $L_K(f, s)$ is the L -function of the rank 2 motive $H_1(E)$ over K .

3. An integral representation. The function $L_K(f, s)$ has an important integral representation in its half-plane of convergence. This involves the Eisenstein series of weight 1 and character ε on $\Gamma_0(p)$, defined by the series

$$E(z, s) = \frac{1}{2} \sum_{\substack{c, d \in \mathbf{Z} \\ c \equiv 0 \pmod{p}}} \frac{\varepsilon(d)}{cz + d} \frac{y^s}{|cz + d|^{2s}}, \quad (3.1)$$

which converges for $\operatorname{Re}(s) > \frac{1}{2}$. Using the Poisson summation formula, one can show that $E(z, s)$ has an analytic continuation to the entire s -plane and satisfies a simple functional equation when s is replaced by $-s$. We denote by $E_N(z, s)$ the same sum, taken over the subset $c \equiv 0 \pmod{Np}$ and $(d, N) = 1$; this also has an analytic continuation to all $s \in \mathbf{C}$.

The holomorphic Eisenstein series $g(z)$ of weight 1 and character ε on $\Gamma_0(p)$ has Fourier expansion

$$g(z) = \frac{h}{2} + \sum_{n \geq 1} R(n) e^{2\pi i n z}, \quad (3.2)$$

and Mellin transform equal to $(2\pi)^{-s} \Gamma(s) \zeta_K(s)$. For consistency, we define $R(0) = h/2$ and $R(n) = 0$ for $n < 0$. By a calculation of the Fourier coefficients of $E(z, s)$, Hecke obtained the identity

$$E(z, 0) = \left(\frac{2\pi}{\sqrt{p}} \right) g(z). \quad (3.3)$$

In terms of the Eisenstein series $E_N(z, s)$, Rankin's integral representation of $L_K(f, s)$ is given by

$$(4\pi)^{-s} \Gamma(s) L_K(f, s) = \iint_{\mathfrak{H}/\Gamma_0(Np)} f(z) \overline{g(z)} E_N(-\bar{z}, s-1) dx dy. \quad (3.4)$$

This gives the analytic continuation of $L_K(f, s)$ to the entire s -plane.

4. The functional equation. If F is a cusp form of weight 2 on $\Gamma_0(M)$ and G is any modular form of weight 2 on $\Gamma_0(M)$, we define the Petersson inner product of F and G by the integral

$$(F, G)_M = 8\pi^2 \iint_{\mathfrak{H}/\Gamma_0(M)} F(z) \overline{G(z)} dx dy. \quad (4.1)$$

We will omit the subscript when $M = N$.

The integral in (3.4) may be viewed as the Petersson product of f with $g(z)E_N(z, \bar{s}-1)$ on $\Gamma_0(Np)$. We will use the trace operations to write this as an inner product on $\Gamma_0(N)$. When N divides M , there is a linear map from forms of weight 2 on $\Gamma_0(M)$ to forms of weight 2 on $\Gamma_0(N)$, defined by

$$\operatorname{tr}_N^M G = \sum_{\gamma \in \Gamma_0(M) \setminus \Gamma_0(N)} G | \gamma. \quad (4.2)$$

If F is a cusp form on $\Gamma_0(N)$ and G is any modular form on $\Gamma_0(M)$, we have the identity

$$(F, G)_M = (F, \text{tr}_N^M G). \quad (4.3)$$

Hence

$$8\pi^2(4\pi)^{-s}\Gamma(s)L_K(f, s) = (f, \text{tr}_N^{Np}\{g(z)E_N(z, \bar{s} - 1)\}).$$

A calculation of the trace of $E_N(z, s)$ from $\Gamma_0(Np)$ to $\Gamma_0(N)$, gives the final integral formula

$$8\pi^2 N^{s-1}(4\pi)^{-s}\Gamma(s)L_K(f, s) = (f, \text{tr}_N^{Np}\{g(z)E(Nz, \bar{s} - 1)\}). \quad (4.4)$$

From (4.4) and a calculation of the Fourier coefficients of the modular form $\text{tr}_N^{Np}\{g(z)E(Nz, \bar{s} - 1)\}$, we can obtain the functional equation for the completed L -function

$$\Lambda_K(f, s) = (Np)^s(2\pi)^{-2s}\Gamma(s)^2 L_K(f, s). \quad (4.5)$$

This functional equation may be written

$$\Lambda_K(f, s) = -\varepsilon(N)\Lambda_K(f, 2 - s). \quad (4.6)$$

We remark that the sign $-\varepsilon(N) = (\frac{-N}{p})$ in this equation depends only on the level N of f and the discriminant $-p$ of K .

5. Special values. We now use formula (4.4) to study $L_K(f, s)$ at the symmetry point $s = 1$ in its functional equation (4.6). If we set $s = 1$ in (4.4) and use (3.3), we find

$$L_K(f, 1) = \left(\frac{1}{\sqrt{p}}\right) (f, \text{tr}_N^{Np}\{g(z)g(Nz)\}). \quad (5.1)$$

If $\varepsilon(N) = +1$ we have $L_K(f, 1) = 0$, as by (4.6) the order of $L_K(f, s)$ at $s = 1$ is odd. In this case the form $G = \text{tr}_N^{Np}\{g(z)g(Nz)\}$ is identically zero. When $\varepsilon(N) = -1$ we can make (5.1) more explicit by a determination of the Fourier coefficients of the form G . If we define $\delta(n) = 1$ if $(n, p) = 1$ and $\delta(n) = 2$ if $n \equiv 0 \pmod{p}$, we find

$$\begin{aligned} G &= \sum_{m \geq 0} b_m e^{2\pi i m z}, \quad \text{where} \\ b_m &= \sum_{n \geq 0} R(mp - nN)\delta(n)R(n) \\ &= hR(m) + \sum_{n=1}^{\lfloor mp/N \rfloor} R(mp - nN)\delta(n)R(n). \end{aligned} \quad (5.2)$$

6. Special derivatives. When $\varepsilon(N) = +1$ we can obtain a similar result for the first derivative $L'_K(f, 1)$. Differentiating (4.4) once with respect to s and setting $s = 1$ gives an expression for $L'_K(f, 1)$ as the inner product of f with the form

$$\phi = \left(\frac{\sqrt{p}}{2\pi}\right) \text{tr}_N^{Np} \left\{ g(z) \left(\frac{\partial}{\partial s}\right) E(Nz, \bar{s} - 1) \Big|_{s=1} \right\}.$$

If H is the holomorphic projection of ϕ , we find that

$$L'_K(f, 1) = \left(\frac{1}{\sqrt{p}} \right) (f, H), \quad (6.1)$$

and the problem remains to determine the Fourier coefficients c_m of the cusp form H of height 2 on $\Gamma_0(N)$.

Define $\sigma(m) = \sum_{d|m} d$ and let $Q_\nu(t)$ be the Legendre function of the second kind with index $\nu \geq 0$. For $m \geq 1$ with $(m, N) = 1$ we find (after considerable calculation) that

$$\begin{aligned} c_m = & - \sum_{n=1}^{[mp/N]} R(mp - nN) \delta(n) \sum_{d|n} \varepsilon(d) \log \left(\frac{n}{d^2} \right) \\ & - \lim_{s \rightarrow 1} \left\{ 2 \sum_{n=1}^{\infty} R(mp + nN) \delta(n) R(n) Q_{s-1} \left(1 + \frac{2nN}{mp} \right) - \frac{12h\sigma(m)}{(N+1)(s-1)} \right\} \\ & - \left(\frac{12h}{N+1} \right) \left\{ \sum_{d|m} d \log \left(\frac{m}{d^2} \right) + \sigma(m) \left[2 + \log \frac{N}{p} + \frac{2 \log N}{N^2 - 1} + 2 \frac{\zeta'}{\zeta}(2)_{-2} \frac{L'}{L}(\varepsilon, 1) \right] \right\} \\ & + hR(m) \left\{ \log \frac{Np}{m} - 2 \log 2\pi + 2 \frac{\Gamma'}{\Gamma}(1) + 2 \frac{L'}{L}(\varepsilon, 1) \right\}. \end{aligned} \quad (6.2)$$

This completes our analytic investigation of $L_K(f, s)$ at $s = 1$.

7. Quaternionic curves. To give an algebraic interpretation of the Fourier coefficients of the modular forms G and H introduced in the previous two sections, we will introduce modular curves of level N and their special points of discriminant $-p$. The curves are associated to orders R of reduced discriminant N in rational quaternion algebras, and the special points correspond to embeddings of \mathcal{O} into the orders R . Since N is prime, there are only two choices for R , up to local conjugacy.

When $\varepsilon(N) = -1$, let B be the definite quaternion algebra over \mathbf{Q} which is ramified at the places ∞ and N . The field K then embeds in B and gives a decomposition $B = K \oplus Kj$, where $j\alpha = \bar{\alpha}j$ for $\alpha \in K$ and $j^2 = -N$. Let R be a maximal order which contains $\mathcal{O} \oplus \mathcal{O}j$; explicitly, R has the form $\{\alpha + \beta j : \alpha \in \mathcal{D}^{-1}, \beta \in \mathcal{D}^{-1}, \alpha \equiv \mu\beta \pmod{\mathcal{O}}\}$, where $\mu^2 \equiv -N \pmod{p}$ and \mathcal{D}^{-1} is the inverse different of \mathcal{O} .

Associated to the quaternion algebra B is a curve Y of genus zero over \mathbf{Q} , defined as a homogeneous conic by the vanishing of the reduced trace and reduced norm. In this case, we have the equation $px^2 + Ny^2 + Npz^2 = 0$ in the projective plane. The group B^*/\mathbf{Q}^* acts on Y by conjugation, and the points $Y(K)$ of Y in K correspond bijectively to the \mathbf{Q} -algebra embeddings of K into B .

Let $\hat{\mathbf{Z}} = \varprojlim \mathbf{Z}/n\mathbf{Z}$ and let $\hat{\mathbf{Q}} = \hat{\mathbf{Z}} \otimes \mathbf{Q}$ be the ring of finite adèles for \mathbf{Q} . Similarly, we define the rings $\hat{\mathcal{O}} = \hat{\mathbf{Z}} \otimes \mathcal{O}$, $\hat{R} = \hat{\mathbf{Z}} \otimes R$, $\hat{K} = \hat{\mathbf{Q}} \otimes K$, and

$\hat{B} = \hat{\mathbf{Q}} \otimes B$. The curve X we will need is defined over \mathbf{Q} as the quotient

$$X = (\hat{R}^* \backslash \hat{B}^* \times Y) / B^*, \quad (7.1)$$

where B^* acts simultaneously on the right of $\hat{R}^* \backslash \hat{B}^*$ and Y .

We note that the double coset space $\hat{R}^* \backslash \hat{B}^* / B^*$ is finite, and corresponds bijectively to the left ideal classes for R . If g_1, \dots, g_n represent the distinct cosets, the associated left ideal $I_i = \hat{R}g_i \cap B$ has right order $R_i = g_i^{-1} \hat{R}g_i \cap B$. The group $\Gamma_i = R_i^* / \mathbf{Z}^*$ embeds as a discrete subgroup of the compact group $(B \otimes \mathbf{R})^* / \mathbf{R}^* \simeq \mathrm{SO}_3(\mathbf{R})$; hence Γ_i is finite, of order w_i . The map taking the class $(\hat{R}g_i, y) \pmod{B^*}$ to the class $y \pmod{\Gamma_i}$ gives an isomorphism

$$X \xrightarrow{\sim} \coprod_{i=1}^n Y / \Gamma_i. \quad (7.2)$$

Hence X is a disjoint union of n curves of genus zero over \mathbf{Q} .

8. The heights of special points. Let z be the point of $Y(K)$ corresponding to our original embedding $i: K \rightarrow B$; then z is fixed by $i(K^*)$. The algebra homomorphism i induces a group homomorphism $\hat{i}: \hat{K}^* \rightarrow \hat{B}^*$ with $i(\hat{K}^*) \cap \hat{R}^* = i(\hat{\mathcal{O}}^*)$. We may therefore define the map

$$\begin{aligned} \hat{\mathcal{O}}^* \backslash \hat{K}^* / K^* &\rightarrow X(K) \\ a &\mapsto x_a = (\hat{i}(a), z), \end{aligned} \quad (8.1)$$

whose image consists of h special points of discriminant $-p$ on X . Let d_K be the divisor $\sum_a (x_a)$ on X over K .

The group $\mathrm{Pic}(X)$ of divisor classes of X is isomorphic to \mathbf{Z}^n . We may fix an isomorphism by taking as \mathbf{Z} -basis $\langle e_1, \dots, e_n \rangle$, where e_i is a divisor of degree 1 on the component Y / Γ_i . We define a height pairing on $\mathrm{Pic}(X)$ with values in \mathbf{Z} by taking the unique bi-additive extension of the pairing

$$\langle e_i, e_i \rangle = w_i, \quad \langle e_i, e_j \rangle = 0, \quad i \neq j. \quad (8.2)$$

The associated bilinear form is positive definite on $\mathrm{Pic}(X) \otimes \mathbf{R}$. Using the geometry of the coset space $\hat{R}^* \backslash \hat{B}^*$, one can define Hecke correspondences T_m on X for all integers $m \geq 1$. These induce endomorphisms of $\mathrm{Pic}(X)$ which are self adjoint with respect to the height pairing (8.2). Our main identity (in the case when $\varepsilon(N) = -1$) is the formula

$$b_m = \langle d_K, T_m d_K \rangle, \quad (8.3)$$

where b_m is the m th Fourier coefficient of the modular form

$$G = \mathrm{tr}_N^{Np} \{g(z)g(Nz)\}$$

entering into formula (5.1) for the value $L_K(f, 1)$. This gives a relation between the heights of special divisors on X and the values of Rankin L -series at $s = 1$.

The proof of (8.3) is accomplished by explicitly computing the height pairings, which match up with formula (5.2) for the Fourier coefficient b_m for $m \geq 1$.

9. Modular curves. When $\varepsilon(N) = +1$ we let B be the split quaternion algebra $M_2(\mathbf{Q})$ over \mathbf{Q} . Let R be the Eichler order $R = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbf{Z}) : c \equiv 0 \pmod{N} \right\}$ in B , which has reduced discriminant N . The curve X is defined over \mathbf{C} as the quotient $X(\mathbf{C}) = \hat{R}^* \backslash \hat{B}^* \times Y(\mathbf{C}) - Y(\mathbf{R}) / B^*$.

In this case Y is isomorphic to \mathbf{P}^1 over \mathbf{Q} , so $Y(\mathbf{C}) - Y(\mathbf{R})$ is analytically isomorphic to the upper and lower half-planes \mathfrak{H}^\pm . The strong approximation theorem shows that the double coset space $\hat{R}^* \backslash \hat{B}^* / B^*$ reduces to a single element, so

$$X(\mathbf{C}) \simeq \mathfrak{H}^\pm / \Gamma \quad (9.2)$$

where $\Gamma = R^* / \mathbf{Z}^*$ embeds as a discrete subgroup of $(B \otimes \mathbf{R})^* / \mathbf{R}^* \simeq \mathrm{PGL}_2(\mathbf{R})$. Since R^* contains an element of determinant -1 , the space $X(\mathbf{C})$ is isomorphic to $\mathfrak{H} / \Gamma_0(N)$.

We may obtain a canonical model for X as a curve over \mathbf{Q} by taking the coarse moduli space for elliptic curves with a subgroup of order N (necessarily cyclic, as N is prime). The open curve X may be compactified by the addition of two cusps ∞ and 0 , which represent the orbits of $\Gamma_0(N)$ on $\mathbf{P}^1(\mathbf{Q})$, corresponding to certain degenerations in the moduli problem. The resulting complete curve over \mathbf{Q} is denoted $X_0(N)$.

10. The heights of Heegner points. Now fix an embedding $i: K \rightarrow B$ taking \mathcal{O} into R ; this is possible by our hypothesis that $\varepsilon(N) = +1$. Let z be the corresponding point in $Y(K) - Y(\mathbf{Q})$. As in §8, we obtain a map

$$\begin{aligned} \hat{\mathcal{O}}^* \backslash \hat{K}^* / K^* &\rightarrow X(\mathbf{C}) \\ a &\mapsto x_a = (i(a), z). \end{aligned} \quad (10.1)$$

The image consists of h Heegner points of discriminant $-p$ on $X_0(N)$. The divisor $D_K = \sum_a (x_a)$ has degree h , and the theory of complex multiplication shows that it is rational over the field K .

Let J denote the Jacobian of $X_0(N)$ over \mathbf{Q} , and let e_K be the class of the divisor $D_K - h(\infty)$ of degree zero in $J(K)$. The theory of canonical heights for abelian varieties, due to Néron and Tate, gives a bi-additive pairing

$$\langle \cdot, \cdot \rangle: J(K) \times J(K) \rightarrow \mathbf{R}$$

which is positive definite on $J(K) \otimes \mathbf{R}$. This pairing can be written as the sum of local terms, over places of K , once disjoint divisors of degree zero representing the paired classes have been chosen.

One can again exploit the geometry of the coset space $\hat{R}^* \backslash \hat{B}^*$ to define Hecke correspondences T_m on $X_0(N)$ for integers $m \geq 1$. These induce endomorphisms of the group $J(K)$ which are self adjoint with respect to the height pairing. Our main identity (in the case when $\varepsilon(N) = +1$) is the formula

$$c_m = \langle e_K, T_m e_K \rangle. \quad (10.2)$$

Here c_m is the m th Fourier coefficient of the cusp form H entering into formula (6.1) for the first derivative $L'_K(f, 1)$. This gives a relation between the heights of Heegner points on $X_0(N)$ and the first derivatives of Rankin L -series at $s = 1$.

The proof of (10.2) is accomplished by explicitly computing the local height pairings for the divisors $D_K - h(\infty)$ and $T_m\{D_K - h(0)\}$ when $(m, N) = 1$. The sum of these terms match up with the formula for c_m obtained in (6.2). Specifically, the first term in (6.2) represents the sum of the local heights at the finite places of K . The second term represents the pairing of D_K and T_mD_K at infinity, using the Green's function for the Poincaré metric on \mathfrak{H} . The third term represents the contribution of the cusps to the local height at infinity, and the final term is a correction factor which occurs when the divisors D_K and T_mD_K are not relatively prime (this situation occurs precisely when $R(m) \neq 0$).

This comparison proves (10.2) for $m \geq 1$ and $(m, N) = 1$. But one can show directly that the series $\sum_{m \geq 1} \langle e_K, T_m e_K \rangle e^{2\pi i m z}$ is a cusp form of weight 2 on $\Gamma_0(N)$, so it follows that the difference of this form and H is an old form. Since there are no forms of weight 2 on $\mathrm{SL}_2(\mathbf{Z})$, we obtain (10.2) for all $m \geq 1$.

11. The limit formulae. Now assume f is a newform. Then f gives a real valued character $T_l \mapsto a_l$ of the Hecke algebra \mathbf{T} acting on modular forms of weight 2 for $\Gamma_0(N)$. The trace formula of Eichler and Selberg shows that \mathbf{T} acts (via the correspondences T_m) on the groups $\mathrm{Pic}(X)$ and $J(K)$ defined in §8 and §10. Since the correspondences commute, and are self adjoint with respect to the definite height pairings, they may be simultaneously diagonalized over \mathbf{R} . When $\varepsilon(N) = -1$ we let $d_{f,K}$ be the f -isotypical component of d_K in $\mathrm{Pic}(X) \otimes \mathbf{R}$, and when $\varepsilon(N) = +1$ we let $e_{f,K}$ be the f -isotypical component of e_K in $J(K) \otimes \mathbf{R}$.

Combining (5.1) with (8.3), we obtain the identity

$$L_K(f, 1) = \frac{(f, f)}{\sqrt{p}} \langle d_{f,K}, d_{f,K} \rangle, \quad (11.1)$$

when $\varepsilon(N) = -1$. Combining (6.1) with (10.2), we obtain the identities

$$L_K(f, 1) = 0, \quad L'_K(f, 1) = \frac{(f, f)}{\sqrt{p}} \langle e_{f,K}, e_{f,K} \rangle, \quad (11.2)$$

when $\varepsilon(N) = +1$.

The formulas (11.1) and (11.2) have many arithmetic corollaries. When f has rational Fourier coefficients, so corresponds to an elliptic curve of conductor N up to isogeny over \mathbf{Q} , they provide strong theoretical evidence for the conjecture of Birch and Swinnerton-Dyer for E over K . For example, if $L(f, 1) = 0$ but $L'(f, 1) \neq 0$, formula (11.2) can be used to show that E has rank ≥ 1 over \mathbf{Q} . In fact, the class $e_{f,K}$ lies in $E(\mathbf{Q}) \otimes \mathbf{Q}$ and is nontrivial for the (infinite number) of quadratic fields K where $L(f \otimes \varepsilon_K, 1) \neq 0$. The conjecture of Birch and Swinnerton-Dyer predicts that the rank is equal to 1 over \mathbf{Q} ; we can at least show that all the classes $e_{f,K}$ all lie on a distinguished line in the vector space $E(\mathbf{Q}) \otimes \mathbf{Q}$. The analogous fact that the classes $d_{f,K}$ lie on a line in $\mathrm{Pic}(X) \otimes \mathbf{Q}$ is clear, as for any eigenform f the space $(\mathrm{Pic}(X) \otimes \mathbf{R})^f$ is 1-dimensional. The relative positions of the classes $e_{f,K}$ and $d_{f,K}$ on these lines as we vary K are determined by the Fourier coefficients of certain cusp forms of weight $\frac{3}{2}$.

BIBLIOGRAPHY

1. B. H. Gross, *Heegner points on $X_0(N)$* , Modular Forms (R. A. Rankin, Ed.) Chichester, Ellis Horwood, 1984, pp. 87–106.
2. B. H. Gross and D. Zagier, *Singular moduli*, J. Crelle **355** (1985), 191–220.
3. ———, *Heegner points and derivatives of L -series*, Invent. Math. **84** (1986), 225–320.
4. B. H. Gross, *Heights and the special values of L -series*, CMS Conf. Proc., no. 7, Amer. Math. Soc., Providence, R. I., 1986.
5. B. H. Gross, W. Kohnen, and D. Zagier, *Heegner points and derivatives of L -series*. II, in preparation.

HARVARD UNIVERSITY, CAMBRIDGE, MASSACHUSETTS 02138, USA

On p -adic Hecke Algebras for GL_2

HARUZO HIDA

In the theory of automorphic functions, holomorphic ones are particularly rich in arithmetic applications, although they are merely a part of the whole space of C^∞ -class modular forms. The analogy between the theory of C^∞ -class modular forms and that of p -adic ones recently becomes a prevalent idea; thus, one may naturally expect the existence of some special class of p -adic modular forms which is as rich in arithmetic structure as the holomorphic counterpart. This paper is intended to present a candidate for such a class of p -adic modular forms in the case of GL_2 over \mathbf{Q} and to explain the reason why we believe that it is the right one.

Our approach is representation theoretic. As in the theory of Jacquet and Langlands in the complex case, we consider the Hecke algebra \mathcal{H} which is the subalgebra of the endomorphism algebra of the space of p -adic modular forms generated topologically by Hecke operators $T(n)$. We then construct a certain idempotent e as a p -adic limit of powers of $T(p)$ in this algebra \mathcal{H} , which gives the projection to the special subspace already mentioned. The Hecke algebra $\mathcal{H}^{\text{ord}} = e\mathcal{H}$ for this special space plays, in our theory for GL_2 , the role of the algebra $\Lambda = \mathbf{Z}_p[[X]]$ in the Iwasawa theory for the cyclotomic \mathbf{Z}_p -extensions, and, naturally, \mathcal{H}^{ord} has a canonical Λ -algebra structure. With each irreducible component of $\text{Spec}(\mathcal{H}^{\text{ord}})$, we can associate a Galois representation into GL_2 with coefficients in the function field of the irreducible component. The infinite algebraic extension corresponding to the kernel of this representation is the object in our theory replacing the cyclotomic \mathbf{Z}_p -extensions. On the other hand, by virtue of Shimura's theory of algebraicity for the special values of zeta functions of modular forms, we are able to construct several p -adic L -functions on the spectrum of each irreducible component of $\text{Spec}(\mathcal{H}^{\text{ord}})$. Thus, what is awaited for further research is the study of the direct arithmetic relation between the Galois representation and the p -adic L -function attached to each irreducible component of $\text{Spec}(\mathcal{H}^{\text{ord}})$.

1. We begin with the definition of Hecke algebras. Let N be a positive integer, and put $\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbf{Z}) \mid c \equiv 0 \pmod{N}, a \equiv d \equiv 1 \pmod{N} \right\}$. For each positive integer k , let $S_k(\Gamma_1(N))$ denote the space of holomorphic cusp forms on

the upper half plane \mathfrak{H} of weight k with respect to $\Gamma_1(N)$. Each element f in $S_k(\Gamma_1(N))$ has the following type of Fourier expansion:

$$f(z) = \sum_{n=1}^{\infty} a(n, f) q^n \quad (q = \exp(2\pi iz), z \in \mathfrak{H})$$

for complex constants $a(n, f)$. By means of this, we may embed $S_k(\Gamma_1(N))$ into the power series ring $\mathbb{C}[[q]]$. One may then give a rational structure on $S_k(\Gamma_1(N))$ through defining the A -rational subspace $S_k(\Gamma_1(N); A)$ for each subalgebra A of \mathbb{C} by $S_k(\Gamma_1(N); A) = S_k(\Gamma_1(N)) \cap A[[q]]$. For each integer l prime to N , by choosing $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ with the congruence: $c \equiv 0 \pmod{N}$ and $d \equiv l \pmod{N}$, we can let l act on $S_k(\Gamma_1(N))$ via

$$f| \langle l \rangle = l^{k-2} f\left(\frac{az+b}{cz+d}\right) (cz+d)^{-k}.$$

Then the Hecke operator $T(n)$ is defined as an endomorphism of $S_k(\Gamma_1(N))$ by

$$a(m, f|T(n)) = \sum_{\substack{l|m, l|n \\ (l, N)=1}} la(mn/l^2, f| \langle l \rangle) \quad (1.1)$$

(see [6, (3.5.12)]). Then $S_k(\Gamma_1(N); A)$ is stable under $T(n)$ and $\langle l \rangle$. The Hecke algebra $\mathcal{H}_k(\Gamma_1(N); \mathbb{Z})$ is by definition the subalgebra of $\mathrm{End}_{\mathbb{C}}(S_k(\Gamma_1(N)))$ generated over \mathbb{Z} by $T(n)$ for all n . For general commutative algebra A , we simply put $\mathcal{H}_k(\Gamma_1(N); A) = \mathcal{H}_k(\Gamma_1(N); \mathbb{Z}) \otimes_{\mathbb{Z}} A$. We let $\overline{\mathbb{Q}}_p$ for each rational prime p denote an algebraic closure of the p -adic field \mathbb{Q}_p . Let $\overline{\mathbb{Q}}$ be the algebraic closure of \mathbb{Q} in \mathbb{C} . We fix for each p an embedding of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$. Thus any algebraic number $a \in \overline{\mathbb{Q}}$ can be considered uniquely as a p -adic number as well as a complex number. Let $|\cdot|_p$ (resp. $|\cdot|$) denote the absolute value of $\overline{\mathbb{Q}}_p$ (resp. \mathbb{C}). We put $S_k(\Gamma_1(N); \overline{\mathbb{Q}}_p) = S_k(\Gamma_1(N); \overline{\mathbb{Q}}) \otimes_{\overline{\mathbb{Q}}} \overline{\mathbb{Q}}_p$ which is naturally a subspace of $\overline{\mathbb{Q}}_p[[q]]$. For each subalgebra A of $\overline{\mathbb{Q}}_p$, put $S_k(\Gamma_1(N); A) = S_k(\Gamma_1(N); \overline{\mathbb{Q}}_p) \cap A[[q]]$.

Now we shall define a pairing

$$\langle \cdot, \cdot \rangle: \mathcal{H}_k(\Gamma_1(N); A) \times S_k(\Gamma_1(N); A) \rightarrow A \quad \text{by } \langle h, f \rangle = a(1, f|h). \quad (1.2a)$$

Then we see from (1.1)

$$\langle T(n), f \rangle = a(n, f) \quad \text{for each positive integer } n. \quad (1.2b)$$

If A is one of the rings \mathbb{Z} , $\overline{\mathbb{Q}}_p$, or \mathbb{C} , then the pairing (1.2a) induces

$$\begin{aligned} \mathrm{Hom}_A(S_k(\Gamma_1(N); A), A) &\cong \mathcal{H}_k(\Gamma_1(N); A), \\ \mathrm{Hom}_A(\mathcal{H}_k(\Gamma_1(N); A), A) &\cong S_k(\Gamma_1(N); A). \end{aligned} \quad (1.3)$$

To each common eigenform f of all Hecke operators $T(n)$ in $S_k(\Gamma_1(N))$, we associate a \mathbb{C} -algebra homomorphism $\lambda: \mathcal{H}_k(\Gamma_1(N); \mathbb{C}) \rightarrow \mathbb{C}$ by $f|h = \lambda(h)f$. Note that

$$\lambda(T(n)) \langle T(1), f \rangle = \langle T(n), f \rangle = a(n, f).$$

Thus if f is nontrivial, then $\langle T(1), f \rangle \neq 0$. Dividing f by $\langle T(1), f \rangle$, we may assume that $\langle T(1), f \rangle = a(1, f) = 1$. Common eigenforms of $T(n)$ under this normalization will be called *normalized* eigenforms. We say that two normalized eigenforms f and g are associated if f and g belong to the same eigenvalues of $T(l)$ except for finitely many primes l . In the class of normalized eigenforms associated with a given f , there exists a unique one contained in $S_k(\Gamma_1(C))$ with the smallest C [5], which is called *primitive*. The level C of the primitive form associated with f is called the *conductor* of f and will be written as $C(f)$. Let A be a subring of $\overline{\mathbf{Q}}_p$ or \mathbf{C} . We denote by B the field $\overline{\mathbf{Q}}_p$ or \mathbf{C} according as $A \subset \overline{\mathbf{Q}}_p$ or $A \subset \mathbf{C}$. If $\lambda: \mathcal{H}_k(\Gamma_1(N); A) \rightarrow B$ is an A -algebra homomorphism, the restriction of λ to $\mathcal{H}_k(\Gamma_1(N); \mathbf{Z})$ has values in $\overline{\mathbf{Q}} \cap A$ since $\mathcal{H}_k(\Gamma_1(N); \mathbf{Z})$ is free of finite rank over \mathbf{Z} by (1.3). Thus we can extend λ to a \mathbf{C} -algebra homomorphism $\lambda_{\mathbf{C}}: \mathcal{H}_k(\Gamma_1(N); \mathbf{C}) \rightarrow \mathbf{C}$. By the duality (1.3), we can find $f_{\lambda} \in S_k(\Gamma_1(N))$ so that $\langle h, f \rangle = \lambda_{\mathbf{C}}(h)$. Especially, we know from (1.2b) that $f_{\lambda}(z) = \sum_{n=1}^{\infty} \lambda(T(n)) q^n$, and then f_{λ} is a normalized eigenform. Therefore, for each subalgebra A of \mathbf{C} or $\overline{\mathbf{Q}}_p$, the correspondence $\lambda \mapsto f_{\lambda}$ induces a bijection

$$\mathrm{Hom}_{A\text{-alg}}(\mathcal{H}_k(\Gamma_1(N); A), B) \cong \{\text{normalized eigenforms in } S_k(\Gamma_1(N))\}. \quad (1.4)$$

Let $\lambda: \mathcal{H}_k(\Gamma_1(N); A) \rightarrow B$ be an A -algebra homomorphism and f be the corresponding normalized eigenform. Then there exists a Dirichlet character $\psi: (\mathbf{Z}/N\mathbf{Z})^{\times} \rightarrow \overline{\mathbf{Q}}$ such that $f| \langle l \rangle = \psi(l) l^{k-2} f$. This character ψ is called the *character* of f or λ .

Now we fix a prime p and a finite extension K/\mathbf{Q}_p in $\overline{\mathbf{Q}}_p$. Let \mathcal{O} be the p -adic integer ring of K . By (1.3), $\mathcal{H}_k(\Gamma_1(N); \mathcal{O})$ is free of finite rank over \mathcal{O} and is then a product of complete local rings R . Let 1_R denote the idempotent of R , and define an idempotent $e_N \in \mathcal{H}_k(\Gamma_1(N); \mathcal{O})$ by the sum of 1_R over the local rings R on which the image of $T(p)$ is invertible. Then

$$\mathcal{H}_k^{\mathrm{ord}}(\Gamma_1(N); \mathcal{O}) = e_N \mathcal{H}_k(\Gamma_1(N); \mathcal{O})$$

is the maximal direct summand of $\mathcal{H}_k(\Gamma_1(N); \mathcal{O})$ on which the image of $T(p)$ is invertible. The idempotent e_N can be explicitly given as a p -adic limit: $e_N = \lim_{n \rightarrow \infty} T(p)^{p^n(p^m-1)} \in \mathcal{H}_k(\Gamma_1(N); \mathcal{O})$ for a suitable positive integer m . If λ is an \mathcal{O} -algebra homomorphism of $\mathcal{H}_k^{\mathrm{ord}}(\Gamma_1(N); \mathcal{O})$ into $\overline{\mathbf{Q}}_p$, we know that $|\lambda(T(p))|_p = 1$. Thus the correspondence (1.4) induces a bijection:

$$\begin{aligned} \mathrm{Hom}_{\mathcal{O}\text{-alg}}(\mathcal{H}_k^{\mathrm{ord}}(\Gamma_1(N); \mathcal{O}), \overline{\mathbf{Q}}_p) \\ \cong \{\text{normalized eigenforms in } S_k(\Gamma_1(N)) \text{ with } |a(p, f)|_p = 1\}. \end{aligned} \quad (1.5)$$

2. Hereafter we suppose that $p \geq 5$ and N is prime to p . If $r \geq s \geq 1$, we have by (1.1) a commutative diagram for all n :

$$\begin{array}{ccc} S_k(\Gamma_1(Np^s); A) & \rightarrow & S_k(\Gamma_1(Np^r); A) \\ \downarrow T(n) & & \downarrow T(n) \\ S_k(\Gamma_1(Np^s); A) & \rightarrow & S_k(\Gamma_1(Np^r); A) \end{array} \quad (2.1)$$

where the horizontal arrows are the natural inclusion. Then the restriction of each

operator in $\mathfrak{h}_k(\Gamma_1(Np^r); A)$ to the subspace $S_k(\Gamma_1(Np^s); A)$ is again contained in $\mathfrak{h}_k(\Gamma_1(Np^s); A)$; thus, we have a surjective A -algebra homomorphism:

$$\mathfrak{h}_k(\Gamma_1(Np^r); A) \rightarrow \mathfrak{h}_k(\Gamma_1(Np^s); A) \quad \text{for each pair } r \geq s \geq 1. \quad (2.2a)$$

Since $T(p)$ is sent to $T(p)$ by (2.2a), the image of e_{Np^r} under (2.2a) coincides with e_{Np^s} , and (2.2a) induces

$$\mathfrak{h}_k^{\text{ord}}(\Gamma_1(Np^r); \mathcal{O}) \rightarrow \mathfrak{h}_k^{\text{ord}}(\Gamma_1(Np^s); \mathcal{O}). \quad (2.2b)$$

We shall take the limits:

$$\mathfrak{h}_k(Np^\infty; \mathcal{O}) = \varprojlim_r \mathfrak{h}_k(\Gamma_1(Np^r); \mathcal{O}), \quad e = \varprojlim_r e_{Np^r} \in \mathfrak{h}_k(Np^\infty; \mathcal{O}),$$

$$\mathfrak{h}_k^{\text{ord}}(Np^\infty; \mathcal{O}) = e\mathfrak{h}_k(Np^\infty; \mathcal{O}) = \varprojlim_r \mathfrak{h}_k^{\text{ord}}(\Gamma_1(Np^r); \mathcal{O}),$$

$$S_k(Np^\infty; \mathcal{O}) = \bigcup_{r=1}^{\infty} S_k(\Gamma_1(Np^r); \mathcal{O}) \quad \text{in } \mathcal{O}[[q]].$$

We define a p -adic norm on $S_k(Np^\infty; \mathcal{O})$ by

$$|f|_p = \sup_n |a(n, f)|_p.$$

Let $\bar{S}_k(Np^\infty; \mathcal{O})$ be the completion of $S_k(Np^\infty; \mathcal{O})$ under this norm. By continuity, $\mathfrak{h}_k(Np^\infty; \mathcal{O})$ naturally acts on $\bar{S}_k(Np^\infty; \mathcal{O})$. Put $\bar{S}_k^{\text{ord}}(Np^\infty; \mathcal{O}) = \bar{S}_k(Np^\infty; \mathcal{O})|e$. Then the duality (1.3) extends to the following (topological and also algebraic) duality [3]:

$$\begin{aligned} \text{Hom}_{\mathcal{O}}(\bar{S}_k(Np^\infty; \mathcal{O}), \mathcal{O}) &\cong \mathfrak{h}_k(Np^\infty; \mathcal{O}), \\ \text{Hom}_{\mathcal{O}}(\bar{S}_k^{\text{ord}}(Np^\infty; \mathcal{O}), \mathcal{O}) &\cong \mathfrak{h}_k^{\text{ord}}(Np^\infty; \mathcal{O}). \end{aligned} \quad (2.3)$$

Therefore we can formulate results for the structure of the space of p -adic modular forms in terms of the Hecke algebras without referring to the space, which we shall do in the rest of the paper.

Define a semigroup \mathcal{X} by $\mathcal{X} = \{l \in \mathbf{Z} | (l, Np) = 1\}$ and a compact group Z by $Z = \varprojlim_r (\mathbf{Z}/Np^r\mathbf{Z})^\times$. We consider \mathcal{X} as a sub-semigroup of Z . We know from (1.1) that the homomorphism of semigroup: $\mathcal{X} \ni l \mapsto \langle l \rangle$ has values in $\mathfrak{h}_k(\Gamma_1(Np^r); \mathcal{O})$ for each $r \geq 1$ and is continuous under the topology induced from Z . By continuity, we extend this homomorphism to Z as a continuous character with values in $\mathfrak{h}_k(\Gamma_1(Np^r); \mathcal{O})$ and hence obtain a continuous character: $Z \rightarrow \mathfrak{h}_k(Np^\infty; \mathcal{O})$. We can naturally identify Z with $\Gamma \times (\mathbf{Z}/Np\mathbf{Z})^\times$, where $\Gamma = \Gamma_1$ and $\Gamma_r = \{x \in \mathbf{Z}_p^\times | x \equiv 1 \pmod{p^r\mathbf{Z}_p}\}$. We shall define continuous group algebras by $\mathcal{A} = \varprojlim_r \mathcal{O}[(\mathbf{Z}/Np^r\mathbf{Z})^\times]$ and $\Lambda = \Lambda_{\mathcal{O}} = \varprojlim_r \mathcal{O}[\Gamma/\Gamma_r]$. By the universality of continuous group algebras, we have canonical \mathcal{O} -algebra homomorphisms: $\mathcal{A} \rightarrow \mathfrak{h}_k(Np^\infty; \mathcal{O})$ and $\Lambda \rightarrow \mathfrak{h}_k(Np^\infty; \mathcal{O})$. The following fact is essentially due to Shimura.

THEOREM 2.1. *For each $k \geq 2$, we have canonical \mathcal{A} -algebra isomorphisms*

$$\mathfrak{h}_k(Np^\infty; \mathcal{O}) \cong \mathfrak{h}_2(Np^\infty; \mathcal{O}) \quad \text{and} \quad \mathfrak{h}_k^{\text{ord}}(Np^\infty; \mathcal{O}) \cong \mathfrak{h}_2^{\text{ord}}(Np^\infty; \mathcal{O}),$$

which take $T(m)$ of weight k to $T(m)$ of weight 2 for all m .

A proof of this fact for $\mathcal{H}_k^{\text{ord}}(Np^\infty; \mathcal{O})$ is given in [2, Theorem 1.1]. By this theorem and (2.3), the space $\bar{S}_k(Np^\infty; \mathcal{O})$ is also independent of $k \geq 2$. We write $\mathcal{H}(N; \mathcal{O})$ for the universal Hecke algebra defined by the theorem and $\mathcal{H}^{\text{ord}}(N; \mathcal{O})$ for $e\mathcal{H}(N; \mathcal{O})$.

We write $[\gamma]$ for the image of $\gamma \in \Gamma$ under the natural embedding of Γ into Λ . When we consider $\gamma \in \Gamma$ as an element of \mathbf{Z}_p , hence, as an element in the coefficient ring \mathcal{O} of Λ , we simply write it as $\gamma \in \Lambda$. We fix a topological generator u of Γ and define an element of Λ by $\omega_{n,r} = [u^{p^{r-1}}] - u^{np^{r-1}}$ ($n, r \in \mathbf{Z}$, $r \geq 1$). Then $\Lambda/\omega_{n,r}\Lambda$ is the maximal quotient of Λ on which Γ_r acts via the character: $\Gamma_r \ni \gamma \mapsto \gamma^n \in \mathcal{O}$. For each $n \geq 0$ and $r \geq 1$, we have by definition a surjective Λ -algebra homomorphism $\rho_{n,r}: \mathcal{H}(N; \mathcal{O}) \rightarrow \mathcal{H}_{n+2}(\Gamma_1(Np^r); \mathcal{O})$ which sends $T(m)$ in $\mathcal{H}(N; \mathcal{O})$ to $T(m)$ in $\mathcal{H}_{n+2}(\Gamma_1(Np^r); \mathcal{O})$. The following result determines the structure of $\mathcal{H}^{\text{ord}}(N; \mathcal{O})$:

THEOREM 2.2. *The Λ -algebra $\mathcal{H}^{\text{ord}}(N; \mathcal{O})$ is free of finite rank over Λ . Moreover, for each pair of integers $n \geq 0$ and $r \geq 1$, $\rho_{n,r}$ induces a Λ -algebra isomorphism*

$$\mathcal{H}^{\text{ord}}(N; \mathcal{O}) \otimes_{\Lambda} \Lambda/\omega_{n,r}\Lambda \cong \mathcal{H}_{n+2}^{\text{ord}}(\Gamma_1(Np^r); \mathcal{O}),$$

which sends $T(m)$ to $T(m)$ for all m .

For the proof, see [1] and [2]. When it is unlikely to cause misunderstanding, we simply write \mathcal{H}^{ord} for $\mathcal{H}^{\text{ord}}(N; \mathcal{O})$. By the universality of the continuous group ring, each continuous character $\xi: \Gamma \rightarrow \bar{\mathbf{Q}}_p$ can be extended to an \mathcal{O} -algebra homomorphism $P_\xi: \Lambda \rightarrow \bar{\mathbf{Q}}_p$. When $\xi(\gamma) = \gamma^n \varepsilon(\gamma)$ for a nonnegative integer n and a finite order character $\varepsilon: \Gamma \rightarrow \bar{\mathbf{Q}}$, we write $P_{n,\varepsilon}$ for P_ξ . Let \mathcal{L} be the quotient field of Λ , and we fix an algebraic closure $\bar{\mathcal{L}}$ of \mathcal{L} . We consider $\bar{\mathbf{Q}}_p$ as a subfield of $\bar{\mathcal{L}}$. Let \mathcal{K} be a finite extension of \mathcal{L} in $\bar{\mathcal{L}}$. We denote by \mathcal{I} the integral closure of Λ in \mathcal{K} . Then \mathcal{I} is known to be free of finite rank over Λ . Put $\mathcal{X}(\mathcal{I}) = \text{Hom}_{\mathcal{O}\text{-alg}}(\mathcal{I}, \bar{\mathbf{Q}}_p)$, and let $\mathcal{X}_{\text{alg}}(\mathcal{I})$ denote the subset of $\mathcal{X}(\mathcal{I})$ consisting of all \mathcal{O} -algebra homomorphisms $\mathcal{I} \rightarrow \bar{\mathbf{Q}}_p$ whose restriction to Λ is of the form $P_{n,\varepsilon}$ for some $0 \leq n \in \mathbf{Z}$ and some finite order character $\varepsilon: \Gamma \rightarrow \bar{\mathbf{Q}}$. When $\mathcal{I} = \Lambda$, $\mathcal{X}(\Lambda) \cong \{x \in \bar{\mathbf{Q}}_p \mid |x|_p < 1\}$ via $P \mapsto P([u]) - 1 \in \bar{\mathbf{Q}}_p$, and in general $\mathcal{X}(\mathcal{I})$ is a covering space of $\mathcal{X}(\Lambda)$. For each $P \in \mathcal{X}_{\text{alg}}(\mathcal{I})$, we define $n(P) \in \mathbf{Z}$ and $\varepsilon_P: \Gamma \rightarrow \bar{\mathbf{Q}}$ by $P|_{\Lambda} = P_{n(P), \varepsilon_P}$. The order of ε_P is written as $p^{r(P)-1}$. Any element F of \mathcal{I} may be regarded as a p -adic analytic function $F: \mathcal{X}(\mathcal{I}) \rightarrow \bar{\mathbf{Q}}_p$ whose value at $P \in \mathcal{X}(\mathcal{I})$ is given by $P(F) \in \bar{\mathbf{Q}}_p$.

Now we consider $\lambda \in \text{Hom}_{\Lambda\text{-alg}}(\mathcal{H}^{\text{ord}}, \bar{\mathcal{L}})$, the combination of λ with the canonical character: $\mathbf{Z} \rightarrow \mathcal{H}^{\text{ord}}$ gives a continuous character of \mathbf{Z} and induces a Dirichlet character $\psi: (\mathbf{Z}/Np\mathbf{Z})^\times \rightarrow \bar{\mathbf{Q}}$ since $\mathbf{Z} = \Gamma \times (\mathbf{Z}/Np\mathbf{Z})^\times$. This character ψ is called the character of λ . Since \mathcal{H}^{ord} is of finite rank over Λ , we can find a finite extension \mathcal{K} such that $\lambda(\mathcal{H}^{\text{ord}}) \subset \mathcal{K}$. The integral closure of \mathcal{O} in \mathcal{K} is of finite rank over \mathcal{O} . When it coincides with \mathcal{O} , we say that λ is *defined* over \mathcal{O} . Changing K by its finite extension, if necessary, we may (and will) assume

$$\lambda \text{ is defined over } \mathcal{O}. \quad (2.4)$$

For each $P \in \mathcal{X}_{\text{alg}}(\mathcal{J})$, we consider the composition $\lambda_P = P \circ \lambda: \mathcal{H}^{\text{ord}} \rightarrow \overline{\mathbf{Q}}_p$. Since $P(\omega_{n(P), r(P)}) = 0$, λ_P factors through $\mathcal{H}^{\text{ord}}/\omega_{n, r} \mathcal{H}^{\text{ord}} \cong \mathcal{H}_{n+2}^{\text{ord}}(\Gamma_1(Np^r); \mathcal{O})$ for $n = n(P)$ and $r = r(P)$. Then by (1.5) there is a unique normalized eigenform f_P in $S_{n(P)+2}(\Gamma_1(Np^{r(P)}))$ with q -expansion: $f_P = \sum_{m=1}^{\infty} \lambda_P(T(m))q^m$. This form is called the *ordinary form belonging to λ at $P \in \mathcal{X}_{\text{alg}}(\mathcal{J})$* . The character of f_P (or λ_P) is given by $\varepsilon_P \psi \omega^{-n(P)}$, where ψ is the character of λ and ω is the Teichmüller character. More generally, we say that a normalized eigenform $f \in S_k(\Gamma_1(Np^r))$ is *ordinary* if $r \geq 1$ and $|a(p, f)|_p = 1$. This notion is intrinsic (i.e. independent of r) because of (2.1). When f_0 is a normalized eigenform in $S_k(\Gamma_1(N))$ ($k \geq 2$) with $|a(p, f_0)|_p = 1$, there is a unique ordinary form $f \in S_k(\Gamma_1(Np))$ such that $a(n, f) = a(n, f_0)$ for all n prime to p . This form f is called the ordinary form *associated* with f_0 . By abusing the language, we say that f_0 belongs to λ if the associated ordinary form f belongs to λ . The following fact is obvious from Theorem 2.2:

(2.5) *Each ordinary form $f \in S_k(\Gamma_1(Np^r))$ belongs to some Λ -algebra homomorphism $\lambda: \mathcal{H}^{\text{ord}}(N; \mathcal{O}) \rightarrow \overline{\mathcal{L}}$.*

Let J be another positive integer prime to p . We say that two Λ -algebra homomorphisms $\lambda: \mathcal{H}^{\text{ord}}(N; \mathcal{O}) \rightarrow \overline{\mathcal{L}}$ and $\mu: \mathcal{H}^{\text{ord}}(J; \mathcal{O}) \rightarrow \overline{\mathcal{L}}$ are *associated* if $\lambda(T(l)) = \mu(T(l))$ except for finitely many primes l .

THEOREM 2.3. *Let $\lambda: \mathcal{H}^{\text{ord}}(N; \mathcal{O}) \rightarrow \overline{\mathcal{L}}$ be a Λ -algebra homomorphism. In the class of all Λ -algebra homomorphisms associated with λ , there is a unique $\lambda_0: \mathcal{H}^{\text{ord}}(C; \mathcal{O}) \rightarrow \overline{\mathcal{L}}$ with the smallest level C .*

The homomorphism λ_0 satisfying the condition of Theorem 2.3 is called *primitive* and the level C of λ_0 will be called the *conductor* of λ .

THEOREM 2.4. *If an ordinary form $f \in S_k(\Gamma_1(Np^r); \mathcal{O})$ has conductor divisible by N and $k \geq 2$, then there is a unique $\lambda \in \text{Hom}_{\Lambda\text{-alg}}(\mathcal{H}^{\text{ord}}(N; \mathcal{O}), \overline{\mathcal{L}})$ defined over \mathcal{O} to which f belongs. The homomorphism λ as above is primitive and has conductor N . Conversely, suppose that $\lambda \in \text{Hom}_{\Lambda\text{-alg}}(\mathcal{H}^{\text{ord}}(N; \mathcal{O}), \mathcal{X})$ is primitive and is of conductor N . Then the conductor of each ordinary form belonging to λ is divisible by N . Moreover, let ψ be the character of λ and ψ_p be its restriction to $(\mathbf{Z}/p\mathbf{Z})^\times$. Then the ordinary form f_P belonging to λ at $P \in \mathcal{X}_{\text{alg}}(\mathcal{J})$ is primitive and is of conductor $Np^{r(P)}$ unless $\varepsilon_P \psi_p \omega^{-n(P)}$ is trivial.*

For the proofs of Theorem 2.3 and Theorem 2.4, see [3].

3. By a result of Deligne, one can attach to $\xi \in \text{Hom}_{\mathcal{O}\text{-alg}}(\mathcal{H}_k(\Gamma_1(Np^r); \mathcal{O}), \overline{\mathbf{Q}}_p)$ a simple representation $\pi(\xi): \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\overline{\mathbf{Q}}_p)$, which is characterized by

(3.1a) $\pi(\xi)$ is unramified outside Np .

(3.1b) Let σ_l be the Frobenius element of $\text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ for each prime l outside Np . Then we have that $\det(1 - \pi(\xi)(\sigma_l)X) = 1 - \xi(T(l))X + l\xi(\langle l \rangle)X^2$.

We shall attach to each Λ -algebra homomorphism $\lambda: \mathcal{H}^{\text{ord}}(N; \mathcal{O}) \rightarrow \mathcal{X}$ a representation $\pi(\lambda): \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{X})$ which interpolates p -adically the Deligne representation $\pi(\lambda_P)$ as in (3.1) for all $P \in \mathcal{X}_{\text{alg}}(\mathcal{J})$. A representation

$\pi: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{K})$ is said to be *continuous* if the following conditions are satisfied:

(3.2a) *there exists an \mathcal{I} -submodule L of \mathcal{K}^2 stable under π such that $L \otimes_{\mathcal{I}} \mathcal{K} = \mathcal{K}^2$ and L is of finite type over \mathcal{I} ;*

(3.2b) *the representation $\pi: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{Aut}_{\mathcal{I}}(L)$ is continuous under the \mathfrak{m} -adic topology for the maximal ideal \mathfrak{m} of \mathcal{I} .*

THEOREM 3.1. *For each Λ -algebra homomorphism $\lambda: \mathcal{H}^{\text{ord}}(N; \mathcal{O}) \rightarrow \mathcal{K}$, there exists a unique representation $\pi(\lambda): \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{K})$ characterized by*

(3.3a) *$\pi(\lambda)$ is continuous and simple;*

(3.3b) *$\pi(\lambda)$ is unramified outside Np ;*

(3.3c) *for the Frobenius element σ_l for each prime l outside Np , we have that $\det(1 - \pi(\lambda)(\sigma_l)X) = 1 - \lambda(T(l))X + l\lambda(\langle l \rangle)X^2$.*

For the proof, see [2]. For each prime ideal \mathcal{P} of \mathcal{I} , let $\mathcal{K}(\mathcal{P})$ be the quotient field of \mathcal{I}/\mathcal{P} . We say that a representation $\tilde{\pi}: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{K}(\mathcal{P}))$ is a residual representation modulo \mathcal{P} of a continuous representation $\pi: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{K})$ if $\tilde{\pi}$ is semisimple and the characteristic polynomial of $\tilde{\pi}(\sigma)$ is the reduction of that of $\pi(\sigma)$ modulo \mathcal{P} for every $\sigma \in \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. We write the residual representation $\tilde{\pi}$ modulo \mathcal{P} as $\pi \bmod \mathcal{P}$. When $\pi: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{K})$ is continuous, we now see the existence of the residual representation $\pi \bmod P$ for each $P \in \mathcal{X}(\mathcal{I})$. Let L be the \mathcal{I} -lattice of \mathcal{K}^2 stable under π as in (3.2a). Since the localization \mathcal{I}_P of \mathcal{I} at P is a discrete valuation ring, $L_P = L \otimes_{\mathcal{I}} \mathcal{I}_P$ is free of rank 2 over \mathcal{I}_P . Therefore π induces a representation $\pi: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \rightarrow \text{GL}_2(\mathcal{I}_P)$. Then $\pi \bmod P$ is given by the semisimplification of $P \circ \pi: \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \xrightarrow{\pi} \text{GL}_2(\mathcal{I}_P) \xrightarrow{P} \text{GL}_2(\overline{\mathbf{Q}}_P)$.

By condition (3.3c), $\pi(\lambda) \bmod P$ is equivalent to $\pi(\lambda_P)$ for all $P \in \mathcal{X}_{\text{alg}}(\mathcal{I})$. This is why we think that the representation $\pi(\lambda)$ is a p -adic interpolation of the Deligne representations $\pi(\lambda_P)$. Since we may assume that $\pi(\lambda_P)$ has values in $\text{GL}_2(\mathcal{O}')$ for a valuation ring \mathcal{O}' finite over \mathcal{O} , $\pi(\lambda_P)$ has its residual representation modulo \mathfrak{p} for the maximal ideal \mathfrak{p} of \mathcal{O}' . The representation $\pi(\lambda_P) \bmod \mathfrak{p}$ can be identified with the residual representation of $\pi(\lambda)$ modulo the maximal ideal \mathfrak{m} of \mathcal{I} . Thus

(3.4) *The residual representation $\pi(\lambda) \bmod \mathfrak{m}$ exists [4].*

It should be noted that Mazur and Wiles [4] have shown that the image of $\pi(\lambda)$ contains $\text{SL}_2(\Lambda_{\mathbf{Z}_p})$ for $\lambda: \mathcal{H}^{\text{ord}}(1; \mathbf{Z}_p) \rightarrow \Lambda_{\mathbf{Z}_p}$ if the image of $\pi(\lambda) \bmod \mathfrak{m}$ contains $\text{SL}_2(\mathbf{Z}/p\mathbf{Z})$. (This condition is verified, for example, if the unique normalized eigenform $\Delta = q\prod_{n=1}^{\infty}(1 - q^n)^{24} \in S_{12}(\text{SL}_2(\mathbf{Z}))$ belongs to λ and $p \neq 11, 23$, and 691 .)

4. Finally, we shall refer to the result for p -adic L -functions. Let J be another positive integer prime to p . For two normalized eigenforms $f \in S_k(\Gamma_1(Np'))$ and $g \in S_k(\Gamma_1(Jp''))$, we shall define algebraic numbers $\alpha_l, \beta_l, \alpha'_l, \beta'_l$ for each prime l

by

$$\sum_{n=1}^{\infty} a(n, f) n^{-s} = \prod_l [(1 - \alpha_l l^{-s})(1 - \beta_l l^{-s})]^{-1},$$

$$\sum_{n=1}^{\infty} a(n, g) n^{-s} = \prod_l [(1 - \alpha'_l l^{-s})(1 - \beta'_l l^{-s})]^{-1}.$$

Let ψ' be the character of f and ψ be the primitive character which induces ψ' . Then we define

$$\mathcal{D}(s, f) = \prod_l [(1 - \bar{\psi}(l) \alpha_l^2 l^{-s})(1 - \bar{\psi}(l) \alpha_l \beta_l l^{-s})(1 - \bar{\psi}(l) \beta_l^2 l^{-s})]^{-1},$$

$$\mathcal{D}(s, f, g) = \prod_l [(1 - \alpha_l \alpha'_l l^{-s})(1 - \alpha_l \beta'_l l^{-s})(1 - \beta_l \alpha'_l l^{-s})(1 - \beta_l \beta'_l l^{-s})]^{-1}.$$

We denote by $\mathcal{D}_p(s, f, g)$ the Euler product obtained from $\mathcal{D}(s, f, g)$ by excluding its Euler p -factor. These L -functions can be continued to meromorphic functions of $s \in \mathbb{C}$, and their algebraicity at certain integer points (for details, see below) is shown by Shimura [7] for the latter series and by Sturm [8] for the former. If f is primitive, we can define a canonical transcendental factor $U_{\infty}(f) \in \mathbb{C}^{\times}$ [2, §10; 3, §4] independent of p such that $\mathcal{D}(k, f)/U_{\infty}(f)$ and $\mathcal{D}(m, f, g)/\pi^{2m-\kappa-k} U_{\infty}(f)$ for $\kappa \leq m < k$ are algebraic numbers ($U_{\infty}(f)$ is independent of g). If $\Omega: \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow \text{GL}_2(\bar{\mathbb{Q}}_p)$ (resp. Ω') denotes the Deligne representation as in (3.1) associated with f (resp. g) and $\tilde{\Omega}$ denotes the contragredient representation of Ω , then, up to finitely many Euler factors, $\mathcal{D}(s, f, g)$ is the zeta function of $\Omega \otimes \Omega': \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow \text{GL}_4(\bar{\mathbb{Q}}_p)$ and $\mathcal{D}(s, f)$ corresponds to the unique three-dimensional subrepresentation of $\Omega \otimes \tilde{\Omega}$.

Let $\lambda: \mathcal{H}^{\text{ord}}(N; \mathcal{O}) \rightarrow \mathcal{I}$ be a primitive Λ -algebra homomorphism defined over \mathcal{O} . Let \mathcal{R} be the unique local ring of $\mathcal{H}^{\text{ord}}(N; \mathcal{O})$ through which λ factors. We suppose

$$\text{Hom}_{\Lambda}(\mathcal{R}, \Lambda) \cong \mathcal{R} \quad \text{as } \mathcal{R}\text{-modules.} \quad (4.1)$$

This condition can be verified when $\pi(\lambda) \bmod \mathfrak{m}$ is simple for the maximal ideal \mathfrak{m} of \mathcal{I} and when the restriction ψ_p of the character ψ of λ to $(\mathbb{Z}/p\mathbb{Z})^{\times}$ is nontrivial and different from ω^{-1} [4, 9]. For example, if the discriminant function $\Delta \in S_{12}(\text{SL}_2(\mathbb{Z}))$ belongs to λ , (4.1) is satisfied if $p \neq 11, 691$. By the multiplication in \mathcal{I} , we have a natural \mathcal{I} -algebra homomorphism: $\mathcal{I} \otimes_{\Lambda} \mathcal{I} \rightarrow \mathcal{I}$. Combining $\lambda \otimes \text{id}: \mathcal{R} \otimes_{\Lambda} \mathcal{I} \rightarrow \mathcal{I} \otimes_{\Lambda} \mathcal{I}$ with this morphism, we can extend λ to an \mathcal{I} -algebra homomorphism of $\mathcal{R} \otimes_{\Lambda} \mathcal{I}$ onto \mathcal{I} , which we denote by $\hat{\lambda}$. We can then decompose $\mathcal{R} \otimes_{\Lambda} \mathcal{I} = \mathcal{K} \oplus \mathcal{A}$ as an algebra direct sum such that the projection of $\mathcal{R} \otimes_{\Lambda} \mathcal{I}$ to the first factor \mathcal{K} coincides with $\hat{\lambda}$. Let $\mathcal{R}(\mathcal{A})$ be the projected image of $\mathcal{R} \otimes_{\Lambda} \mathcal{I}$ in \mathcal{A} . Then we have the diagonal map $\delta: \mathcal{R} \otimes_{\Lambda} \mathcal{I} \rightarrow \mathcal{I} \oplus \mathcal{R}(\mathcal{A})$. Under the assumption (4.1), we can find $0 \neq H \in \mathcal{I}$ so that $\text{Coker}(\delta) \cong \mathcal{I}/H\mathcal{I}$ as \mathcal{I} -modules. The function $H: \mathcal{X}(\mathcal{I}) \rightarrow \bar{\mathbb{Q}}_p$ is determined up to unit factors in \mathcal{I} . Let ψ be the character of λ , and write the p -part of ψ as ω^a for $0 \leq a < p-1$.

THEOREM 4.1. *Suppose (4.1) and that $a \neq 0$ or $N = 1$. Let f_p be the ordinary form belonging to λ at $P \in \mathcal{X}_{\text{alg}}(\mathcal{J})$. Then we can find $U_p(f_p) \in \overline{\mathbf{Q}}_p$ with $|U_p(f_p)|_p = 1$ such that*

$$H(P) = \mathcal{D}(n(P) + 2, f_p)/U_\infty(f_p)U_p(f_p).$$

A part of this result is given in [2] and [3], and the proof in the general case will be given in a forthcoming paper.

Now let $\mu: \mathcal{H}^{\text{ord}}(J; \mathcal{O}) \rightarrow \mathcal{J}$ be another primitive Λ -algebra homomorphism defined over \mathcal{O} . For each $Q \in \mathcal{X}_{\text{alg}}(\mathcal{J})$, we denote by g_Q the ordinary form belonging to μ at Q . Put, for each finite order character $\chi: \mathbf{Z}_p^\times \rightarrow \overline{\mathbf{Q}}$ and complex conjugation ρ ,

$$f|\chi = \sum_{n=1}^{\infty} \chi(n)a(n, f)q^n, \quad f^\rho = \sum_{n=1}^{\infty} a(n, f)^\rho q^n \quad (f \in \mathbf{C}[[q]]).$$

If f is a modular form, $f|\chi$ and f^ρ are again modular forms. We write the restriction of the character of μ (resp. λ) to $(\mathbf{Z}/p\mathbf{Z})^\times$ (resp. $(\mathbf{Z}/N\mathbf{Z})^\times$) as ω^b for $0 \leq b < p-1$ (resp. ψ_0). Put, for each triple $(P, Q, R) \in \mathcal{X}_{\text{alg}}(\mathcal{J})^2 \times \mathcal{X}_{\text{alg}}(\Lambda)$,

$$\begin{aligned} k &= n(P) + 2, & m &= n(R), & j &= 2 + 2m + n(Q), \\ v &= r(Q) + 2r(R), & r &= r(P), \end{aligned}$$

and

$$\begin{aligned} t(P, Q, R) &= 2^{-k-j}(\sqrt{-1})^{j-k} \phi(Np) \phi(N/C)^{-1} C^{-1} N^{-1-k/2} J^{m+k/2} \\ &\quad \times [N, J]^{1-m+(j-k)/2} p^{-r+(vj-rk)/2} \Gamma(j-m) \Gamma(m+1) \Gamma(k)^{-1}, \end{aligned}$$

where $[N, J]$ is the least common multiple of N and J , C is the conductor of ψ_0 , ϕ is the Euler function, and $\Gamma(x)$ is the gamma function. For each primitive form $f \in S_k(\Gamma_1(M))$, we define a constant $W(f)$ with $|W(f)| = 1$ by the first Fourier coefficient of $M^{-k/2}f(-1/Mz)z^{-k}$. Let $\mathcal{J} \hat{\otimes}_{\mathcal{O}} \mathcal{J} \hat{\otimes}_{\mathcal{O}} \Lambda$ (resp. $\mathcal{J} \hat{\otimes}_{\mathcal{O}} \mathcal{J}$) be the p -adic completion of $\mathcal{J} \otimes_{\mathcal{O}} \mathcal{J} \otimes_{\mathcal{O}} \Lambda$ (resp. $\mathcal{J} \otimes_{\mathcal{O}} \mathcal{J}$). Any element $F \in \mathcal{J} \hat{\otimes}_{\mathcal{O}} \mathcal{J} \hat{\otimes}_{\mathcal{O}} \Lambda$ can be regarded as a p -adic analytic function $F: \mathcal{X}(\mathcal{J}) \times \mathcal{X}(\mathcal{J}) \times \mathcal{X}(\Lambda) \rightarrow \overline{\mathbf{Q}}_p$.

THEOREM 4.2. *Let the assumption and the notation be as in Theorem 4.1 for λ . Then for each integer $0 \leq c < p-1$, we have a unique p -adic L -function $D \in \mathcal{J} \hat{\otimes}_{\mathcal{O}} \mathcal{J} \hat{\otimes}_{\mathcal{O}} \Lambda$ such that for $(P, Q, R) \in \mathcal{X}_{\text{alg}}(\mathcal{J}) \times \mathcal{X}_{\text{alg}}(\mathcal{J}) \times \mathcal{X}_{\text{alg}}(\Lambda)$,*

$$D(P, Q, R) = t(P, Q, R) a(p, f_p)^{r-v} W(f_p)^{-1} W(g_Q|_{\varepsilon_R}) \frac{\mathcal{D}_p(j-m, f_p, g_Q^{\rho} |\omega^{-c} \varepsilon_R^{-1})}{\pi^{j-k} U_\infty(f_p) U_p(f_p)}$$

if $0 \leq n(R) < n(P) - n(Q)$, $\omega^{-c} \varepsilon_R^{-1} \neq \varepsilon_Q \omega^b$, $\varepsilon_R \omega^c \neq 1$, and $\omega^a \varepsilon_P \neq 1$.

The condition for characters ε_P , ε_Q , ε_R , ω^a , ω^b , and ω^c is not necessary for the evaluation of $D(P, Q, R)$, but without this assumption, a new type of Euler p -factor appears in the right-hand side of the formula and the result becomes a little more complicated. If one wants to interpolate the value \mathcal{D} without the modification at the Euler p -factor, there appears a singularity for the p -adic

L -function obtained:

THEOREM 4.3. Put $t(P, Q) = t(P, Q, P_{0,\iota})$ for $(P, Q) \in \mathcal{X}_{\text{alg}}(\mathcal{J})^2$, where ι is the identity character of Γ . Let the assumption and the notation be as in Theorem 4.1 for λ . Let $X, Y \in \mathcal{J} \hat{\otimes}_{\theta} \mathcal{J}$ be the functions on $\mathcal{X}(\mathcal{J})^2$ defined by $X(P, Q) = u^{n(P)} \varepsilon_P(u) - 1$ and $Y(P, Q) = u^{n(Q)} \varepsilon_Q(u) - 1$ on $\mathcal{X}_{\text{alg}}(\mathcal{J})^2$. Then we have a unique p -adic L -function Δ on $\mathcal{X}(\mathcal{J})^2$ such that

- (i) $(X - Y)\Delta \in \mathcal{J} \hat{\otimes}_{\theta} \mathcal{J}$,
- (ii)

$$\begin{aligned} \Delta(P, Q) &= t(P, Q) a(p, f_P)^{r(P)-r(Q)-2} W(f_P)^{-1} W(f_Q) \\ &\quad \times \frac{\mathcal{D}(n(Q) + 2, f_P, f_Q^p)}{\pi^{n(Q)-n(P)} U_{\infty}(f_P) U_p(f_P)} \end{aligned}$$

if $n(P) > n(Q) \geq 0$ and $\varepsilon_P \omega^a \neq 1$, $\varepsilon_Q \omega^a \neq 1$,

(iii) $((X - Y)\Delta)(P, Q)|_{P=Q} = (1 + Y(P))(p - 1)p^{-1}(\log(u))H(P)$, where $H \in \mathcal{J}$ is as in Theorem 4.1 and $\log(u)$ is the p -adic logarithm of the generator $u \in \Gamma$.

For the proof, see [3, Theorem 10.1]. The above formula (iii) is the p -adic analogue of the well known residue formula of the complex case:

$$\text{Res}_{s=k} \mathcal{D}(s, f, f^p) = \Gamma(k)^{-1} N^{-2} 2^{2k-1} \pi^{k+1} \phi(N)(f, f)$$

for each normalized eigenform $f \in S_k(\Gamma_1(N))$, where

$$(f, f) = \int_{\Gamma_0(N) \backslash H} |f(z)|^2 y^{k-2} dx dy$$

(see [7, (2.5)]).

REFERENCES

1. H. Hida, *Iwasawa modules attached to congruences of cusp forms*, Ann. Sci. École Norm. Sup. (4) **19** (1986), 231–273.
2. ———, *Galois representations into $GL_2(\mathbb{Z}_p[[X]])$ attached to ordinary cusp forms*, Invent. Math. **85** (1986), 545–613.
3. ———, *A p -adic measure attached to the zeta functions associated with two elliptic modular forms II*, Ann. Inst. Fourier (Grenoble) (to appear).
4. B. Mazur and A. Wiles, *On p -adic analytic families of Galois representations*, Preprint.
5. T. Miyake, *On automorphic forms on GL_2 and Hecke operators*, Ann. of Math. **94** (1971), 174–189.
6. G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Kanô Memorial Lectures, No. 1, Publications of the Mathematical Society of Japan, No. 11, Iwanami Shoten, Tokyo; Princeton University Press, Princeton, N. J., 1971.
7. ———, *The special values of the zeta functions associated with cusp forms*, Comm. Pure Appl. Math. **29** (1976), 783–804.
8. J. Sturm, *Special values of zeta functions, and Eisenstein series of half integral weight*, Amer. J. Math. **102** (1980), 219–240.
9. J. Tilouine, *Un sous groupe p -divisible de la jacobienne de $X_1(Np^r)$ comme module sur l'algèbre de Hecke*, Preprint.

Spectral Theory of Automorphic Functions and Recent Developments in Analytic Number Theory

HENRYK IWANIEC

Introduction. Number theory and modular forms have been intimately related from the very beginnings of the latter. Very well known are connections with higher arithmetic and algebraic number theory. In the last decade there have been considerable developments in modular forms in relation to analytic number theory. Its foundation rests on the spectral theory of automorphic functions due to H. Maass [26] and A. Selberg [34]. Initially the spectral theory served to give extremely sharp estimates for certain exponential sums of primary importance for problems in analytic number theory. This area of research was given the nickname “Kloostermania.” Today analytic number theory is paying back. The connection in question has been reversed to that the machinery of analytic number theory is working to improve the spectral theory of automorphic functions, at least in the most interesting case of congruence groups. This indispensable interrelation is the subject of my talk.

We shall also mention some applications to particular problems. My favorite one is the mean-value theorem for primes in arithmetic progressions to large moduli. In 1965 E. Bombieri [1] and A. I. Vinogradov [40] proved powerful results which serve as a substitute for the Riemann hypothesis for Dirichlet’s L -series, especially in reference to sieve methods. In a series of joint papers with E. Bombieri, E. Fouvry, and J. Friedlander (see the references in [2]) we attempted to extend the Bombieri and Vinogradov results beyond the capability of the Riemann hypothesis. The dream just came through in [2] where we prove that for any $a \neq 0$, $A > 0$, and $x \geq 3$ one has,

$$\sum_{\substack{(q,a)=1 \\ q \leq \sqrt{x}(\log x)^A}} \left| \pi(x; q, a) - \frac{1}{\phi(q)} \pi(x) \right| \ll \frac{x}{(\log x)^3} (\log \log x)^B \quad (0.1)$$

with some absolute constant B and the constant implied in \ll depending at most on a and A .

Supported by the Institute for Advanced Study at Princeton, N. J., and partly by Stanford University, CA.

Of numerous other applications let us mention three: (1) an estimate for the Riemann zeta function [6] with an application to upper bounds for the differences between consecutive primes [20]; (2) Fouvry's Brun-Titchmarsh theorem [10] with an application to the last Fermat Theorem; and (3) the result of Gupta and Murty [13] (later refined by Heath-Brown [14]) on the Artin conjecture for primitive roots.

1. Spectral theory. Cuspidal delicacy. Let Γ be a discontinuous group of motions of the hyperbolic plane H with finite covolume $|F| = \text{vol}(\Gamma \backslash H) < \infty$ and of noncompact type. Thus Γ is a finitely generated subgroup of $\text{PSL}(2, \mathbf{R})$ with a finite number of inequivalent cusps, say $h \geq 1$. A function $f: H \rightarrow \mathbf{C}$ is Γ -automorphic if $f(\gamma z) = f(z)$ for $\gamma \in \Gamma$ and $z \in H$. Let $L^2(\Gamma \backslash H)$ denote the space of Γ -automorphic functions which are square-integrable on the quotient $F = \Gamma \backslash H$ with respect to the invariant measure $d\mu z = y^{-2} dx dy$, and let

$$\Delta = y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$$

denote the Laplace-Beltrami operator for the Poincaré metric

$$ds^2 = y^{-2}(dx^2 + dy^2).$$

The most fundamental problems in the theory of automorphic functions are concerned with the decomposition of $L^2(\Gamma \backslash H)$ into eigenspaces of Δ . Since Γ is not compact, Δ has a continuous spectrum which covers the half-line $[\frac{1}{4}, \infty)$ uniformly with finite multiplicity equal to h . The eigenfunctions of the continuous spectrum are relatively well understood; they are the Eisenstein series $E_{\mathfrak{a}}(z, s)$ on $s = \frac{1}{2} + it$, $t \in \mathbf{R}$, where \mathfrak{a} ranges over a complete set of Γ -inequivalent cusps. Let $\Gamma_{\mathfrak{a}} = \{\gamma \in \Gamma; \gamma \mathfrak{a} = \mathfrak{a}\}$ be the stability group of \mathfrak{a} . Choose $\sigma_{\mathfrak{a}} \in \text{PSL}(2, \mathbf{R})$ (once and for all) such that $\sigma_{\mathfrak{a}}^{-1} \Gamma_{\mathfrak{a}} \sigma_{\mathfrak{a}}$ is the cyclic group of translations $z \rightarrow z + b$ by integers $b \in \mathbf{Z}$. Then $E_{\mathfrak{a}}(z, s)$ is defined by

$$E_{\mathfrak{a}}(z, s) = \sum_{\gamma \in \Gamma_{\mathfrak{a}} \backslash \Gamma} (\text{Im } \sigma_{\mathfrak{a}} \gamma z)^s,$$

where the series converges absolutely in $\text{Re } s > 1$. The problem of meromorphic continuation to the whole complex s -plane is intricate in general; see [15] and the references therein. For congruence groups one can easily establish the continuation by examining the Fourier expansion

$$E_{\mathfrak{a}}(\sigma_{\mathfrak{b}} z, s) = \delta_{\mathfrak{a}\mathfrak{b}} y^s + \varphi_{\mathfrak{a}\mathfrak{b}}(s) y^{1-s} + y^{1/2} \sum_{n \neq 0} \varphi_{\mathfrak{a}\mathfrak{b}n}(s) K_{s-1/2}(2\pi |n|y) e(nx)$$

where the coefficients $\varphi_{\mathfrak{a}\mathfrak{b}}(s)$ are given by Dirichlet's series

$$\varphi_{\mathfrak{a}\mathfrak{b}}(s) = \pi^{1/2} \frac{\Gamma(s - \frac{1}{2})}{\Gamma(s)} \sum_{c > 0} \omega(c) c^{-2s}$$

with $\omega(c) = \#\left\{ \begin{pmatrix} * & * \\ c & * \end{pmatrix} \in \Gamma_{\mathfrak{a}} \backslash \sigma_{\mathfrak{a}}^{-1} \Gamma \sigma_{\mathfrak{b}} / \Gamma_{\mathfrak{b}} \right\}$ and $\varphi_{\mathfrak{a}\mathfrak{b}n}(s)$ expressed similarly. When Γ is a congruence group, $\varphi_{\mathfrak{a}\mathfrak{b}}(s)$ is closely identified with the zeta-function of a number field, so $\varphi_{\mathfrak{a}\mathfrak{b}}(s)$ is meromorphic of order 1.

The theory of the continuous spectrum depends intimately on the analytic structure of $\varphi_{ab}(s)$. The constant terms $\varphi_{ab}(s)$ form the so-called “scattering matrix” $\Phi(s) = [\varphi_{ab}(s)]$, which is unitary on $\operatorname{Re} s = \frac{1}{2}$ and satisfies the functional equation $\Phi(s)\Phi(1-s) = I$. Hence the column-vector of the Eisenstein series $\mathcal{E}(z, s) = {}^t[E_a(z, s)]$ inherits the functional equation

$$\mathcal{E}(z, s) = \Phi(s)\mathcal{E}(z, 1-s).$$

In the half-plane $\operatorname{Re} s > \frac{1}{2}$ the poles of $\Phi(s)$ and $\mathcal{E}(z, s)$ coincide; they are all simple and real. The point $s = s_0 = 1$ is always a pole and the residuum is constant;

$$\operatorname{res}_{s=1} E_a(z, s) = |F|^{-1}.$$

For congruence groups there are no more poles in $\operatorname{Re} s > \frac{1}{2}$, but in general there can be a finite number of other poles s_j , say, with $\frac{1}{2} < s_j < 1$. The residua $u_j(z)$ at $s = s_j$ are eigenfunctions of Δ with eigenvalue $\lambda_j = s_j(1-s_j)$; they are square-integrable (in contrast to the Eisenstein series) and belong to the subspace of the point spectrum of Δ . The remaining part of the point spectrum (possibly empty) is spanned by Maass cusp forms, i.e., by Γ -automorphic functions u_j such that $(\Delta + \lambda_j)u_j = 0$, $\lambda_j = s_j(1-s_j)$, and that the zero Fourier coefficient of u_j at any cusp vanishes.

The existence and construction of cusp forms for Γ has been a frustrating problem for many years. It was a common opinion (based on observation of arithmetic groups) that the space of the point spectrum is always infinite-dimensional (Roelcke’s problem). Somewhat more optimistic doctrines asserted that the point spectrum dominates over the continuous one in a certain quantitative sense. To explain this let us consider the counting function of Δ -eigenvalues

$$N_\Gamma(T) = \#\{j; \lambda_j = s_j(1-s_j) \text{ with } s_j = \tfrac{1}{2} + it_j, |t_j| \leq T\}$$

together with

$$M_\Gamma(T) = \frac{1}{4\pi} \int_{-T}^T \frac{-\varphi'}{\varphi} \left(\frac{1}{2} + it \right) dt,$$

where $\varphi(s)$ is the determinant of the scattering matrix $\Phi(s)$. The second quantity accounts for the continuous spectrum. Notice that $M_\Gamma(T)$ is positive since it is equal to the number of poles of $\Phi(s)$, $s = \sigma + it$, in $\sigma < \frac{1}{2}$, $|t| < T$ (up to some bounded quantity). It is not difficult to deduce from the Selberg’s trace formula (see [39]) that

$$N_\Gamma(T) + M_\Gamma(T) = |F|T^2/4\pi + O(T \log T).$$

Had $\Phi(s)$ been meromorphic of order 1 we could have $M_\Gamma(T) \ll T^{1+\varepsilon}$ and conclude

$$N_\Gamma(T) \sim |F|T^2/4\pi \quad \text{as } T \rightarrow \infty,$$

which is called Weyl’s law. From what we said earlier it follows that Weyl’s law is true for congruence groups. This statement is due to A. Selberg [34]. He also conjectured that for general Γ there is $\eta = \eta(\Gamma) > 0$ such that $M_\Gamma(T) = O(T^{2-\eta})$

with the effect that every Γ is essentially “cuspidal.” Paradoxically, it is not even known whether every Γ has a cusp form.

To everybody’s surprise R. S. Phillips and P. Sarnak have recently revealed a very different truth [28, 29, 30]. They studied the behavior of a cusp form $u_j(z)$ of the Hecke group $\Gamma = \Gamma_0(q)$ under certain quasiconformal deformations showing that $u_j(z)$ is “destroyed” (a constant term appears) provided some L -function does not vanish at the special point $s = s_j$. The L -function in question is the Rankin-Selberg convolution $L(u_j \otimes f, s)$ attached to $u_j(z)$ and to a holomorphic cusp form $f(z)$, say, of weight 4 for $\Gamma_0(q)$ (f exists for large q). The holomorphic form $f(z)$ is the one which Phillips and Sarnak used to create quadratic differentials for perturbing the metric ds^2 . Let the Fourier expansions of u_j and f at the cusp ∞ be given by

$$u_j(z) = y^{1/2} \sum_{n \neq 0} \rho_j(n) K_{s_j-1/2}(2\pi|n|y) e(nx) \quad (1.1)$$

and

$$f(z) = \sum_{n=1}^{\infty} a_n e(nz).$$

We normalize by setting $\lambda_j(n) = \rho_j(n) \Gamma(s_j) |F|^{1/2}$ and $\alpha_n = a_n n^{-3/2}$. Then the Rankin-Selberg L -function is defined by

$$L(u_j \otimes f, s) = \zeta_q(2s) \sum_{n=1}^{\infty} \alpha_n \lambda_j(n) n^{-s},$$

where $\zeta_q(s) = \prod_{p \nmid q} (1 - p^{-s})^{-1}$. It satisfies the functional equation $s \leftrightarrow 1 - s$, so $\operatorname{Re} s = \frac{1}{2}$ is the critical line. The question of which $L(u_j \otimes f, s)$ do not vanish at the special point $s = s_j$ is an open and difficult one. J.-M. Deshouillers and H. Iwaniec [8] have proved that $L(u_j \otimes f, s_j) \neq 0$ for at least $T(\log T)^{-2}$ points $s_j = \frac{1}{2} + it_j$ of height $|t_j| < T$ counted with multiplicity. From this Phillips and Sarnak concluded that infinitely many cusp forms for $\Gamma_0(q)$ disappear under a generic deformation, and generically, of course, none are created. If we accept such standard conjectures as the Lindelöf hypothesis $L(u_j \otimes f, s_j) \ll |s_j|^\varepsilon$ and the bound $m_j \ll |s_j|^\varepsilon$ for the multiplicity m_j of $\lambda_j = s_j(1 - s_j)$, then we conclude [9]

THEOREM (DESHOULLERS-IWANIEC-PHILLIPS-SARNAK). *Generically for $\Gamma \in \mathcal{T}(\Gamma_0(q))$ we have $M_\Gamma(T) \gg T^{2-\varepsilon}$.*

Here $\mathcal{T}(\Gamma_0(q))$ is the Teichmüller space of $\Gamma_0(q)$ and a property is called “generic” if it holds for all members of $\mathcal{T}(\Gamma_0(q))$ except for a set of first Baire category.

In view of the above results it is plausible to expect even more [29].

CONJECTURE (PHILLIPS AND SARNAK). The generic $\Gamma \in \mathcal{T}(\Gamma_0(q))$ has only a finite number of linearly independent cusp forms.

The conjecture of Phillips and Sarnak has tremendous impact on the global picture of the spectral theory of automorphic functions. Now one knows why a

general method of constructing cusp forms was not found. We realize how fortunate are the congruence groups in having so many cusp forms. That congruence groups are essentially cuspidal makes the spectral theory rich and particularly interesting from the point of view of arithmetic.

2. Selberg's trace formula versus Kloosterman sums formula. The celebrated formula of Selberg [35] connects the spectrum of the Laplace operator on $\Gamma \backslash H$ with the length of closed geodesics of $\Gamma \backslash H$;

$$\sum_{j \geq 0} h(t_j) = \frac{|F|}{4\pi} \int_{-\infty}^{\infty} t \tanh(\pi t) h(t) dt + \sum_{\{P\}} \frac{(\Lambda P) \hat{h}(\log NP)}{(NP)^{1/2} - (NP)^{-1/2}} + \cdots, \quad (2.1)$$

where \hat{h} is the Fourier transform of h . The sequence of numbers $\{\log NP\}$, where $\{P\}$ ranges over the conjugacy classes of hyperbolic elements, is called the length spectrum. These two spectra follow dual courses. Each one determines the other (for strictly hyperbolic groups). A counterpart of Weyl's law is the "prime geodesic theorem" (see, for example, P. Sarnak's thesis [32]),

$$\psi_{\Gamma}(x) = \sum_{NP \leq x} \log NP = \sum_{1/2 < s_j < 1} s_j^{-1} x^{s_j} + O(x^{v+\varepsilon}), \quad \frac{1}{2} \leq v < 1,$$

which resembles the familiar prime number theorem. This analogue becomes more suggestive when the trace formula is specified to the spectral decomposition of Z'_{Γ}/Z_{Γ} , where Z_{Γ} is the Selberg zeta-function.

Recently N. V. Kuznetsov [24] and R. W. Bruggeman [3] derived new formulas that connect the spectrum of Δ and the Fourier coefficients of automorphic forms on one side with Kloosterman sums on the other side. Let $\Gamma = \Gamma_0(q)$. Choose an orthonormal basis $\{u_j\}$ of the space $L_0^2(\Gamma \backslash H)$ of cusp forms whose Fourier expansion at the cusp ∞ is given by (1.1). For simplicity we display only the cuspidal fragment of Kuznetsov's formula:

$$\frac{1}{|F|} \sum_{j \geq 1} f(t_j) \lambda_j(m) \bar{\lambda}_j(n) + \cdots = \sum_{c \equiv 0 \pmod{q}} c^{-1} S(m, n; c) \tilde{f}\left(\frac{4\pi\sqrt{mn}}{c}\right), \quad (2.2)$$

where \tilde{f} is a Bessel transform of f . One should realize that the above relation is a direct extension of the Petersson formula for coefficients of the holomorphic Poincaré series $P_m(z, k)$ to the nonholomorphic ones. The nonholomorphic Poincaré series $U_m(z, s)$ were first introduced by A. Selberg [36], who also established the main ingredients for the proof of (2.2), namely, the meromorphic continuation of $U_m(z, s)$ and formulas for the inner product $\langle U_m, u_j \rangle$. By means of these formulas Selberg was able to conclude the meromorphic continuation of the zeta-function

$$\mathfrak{z}_{mn}(s) = \sum_{c > 0} c^{-2s} S(m, n; c) \quad (2.3)$$

and the distribution of its poles. Had he controlled the growth of $\mathfrak{z}_{mn}(s)$ in the whole s -plane, Selberg could have essentially proved (2.2) in his way. This way became possible only later due to the work of D. Goldfeld and P. Sarnak [11].

An alternative way of proving (2.2) was suggested by D. Zagier. Much as in the derivation of the trace formula (2.1) one looks at the spectral decomposition of the automorphic kernel of an invariant integral operator

$$k(z, w) = \sum_{\gamma \in \Gamma} k(d(z, \gamma w)) = \sum_j k^*(t_j) u_j(z) \bar{u}_j(w) + \cdots,$$

but, instead of integrating over the diagonal $z = w$ to obtain the spectral trace, one picks up the double Fourier coefficients. The calculations on the other side are also different, for instead of arranging the summation over $\gamma \in \Gamma$ into conjugacy classes one uses the double coset decomposition à la Bruhat. Consequently, by Poisson's summation, a sum of Kloosterman sums emerges instead of the sum of lengths of closed geodesics. The argument does not end yet because in the resulting formula the test function f and its transform \tilde{f} are subject to quite severe restrictions. In order to release these restrictions one needs a completeness theorem for Bessel functions, which is solely due to Kuznetsov.

It is not wise but it is possible to deduce the Selberg trace formula (2.1) from the Kloosterman sums formula (2.2) for the modular group; see [25], where some initial steps are given. At some point one must use the Dirichlet class number formula $2h(d) \log \varepsilon_d = \sqrt{d} L(1, \chi_d)$ as well as the interpretation of the length of closed geodesics by $\log NP = 2 \log \varepsilon_d$, each one having multiplicity $h(d)$. Here $h(d)$ is the class number of primitive quadratic forms $ax^2 + bxy + cy^2$ with discriminant $d = b^2 - 4ac > 0$, $d \neq \square$, and $\varepsilon_d = \frac{1}{2}(m + n\sqrt{d}) > 1$ with $m^2 - dn^2 = 4$ is the fundamental unit (for details see [32]). In view of the above discussion we may say that the two different partitions of Γ into conjugacy classes and into double cosets respectively are linked up with the classical Dirichlet class number formula. This observation deserves further penetration.

To compare the strengths of (2.1) and (2.2) let us look at the prime geodesic theorem for the modular group. The Selberg trace formula or, what amounts to the same thing, the Selberg zeta-function theory yields the "explicit formula"

$$\psi_\Gamma(x) = x + \sum_{s_j = \frac{1}{2} + it_j, |t_j| \leq T} s_j^{-1} x^{s_j} + O(T^{-1} x \log^2 x)$$

with $1 \leq T \leq \sqrt{x}/\log x$, whence by Weyl's law

$$\psi_\Gamma(x) = x + O(Tx^{1/2} + T^{-1}x \log^2 x) = x + O(x^{3/4} \log x)$$

on taking the optimal $T = x^{1/4} \log x$. If one wishes to reduce the exponent $v = \frac{3}{4}$ one must take into account a cancellation of the terms $s_j^{-1} x^{s_j}$. Such a task cannot be realized within the range of formal Fourier analysis. Similarly, if we use the Kloosterman sums formula (2.1) together with Weil's bound (3.1) we may end up with the same exponent $v = \frac{3}{4}$; see [18]. In order to beat $v = \frac{3}{4}$ one must take into account a cancellation of terms in sums of the type

$$\sigma(n, C, D) = \sum_{c \leq C} c^{-1} S(n, n; c) e(nD/c). \quad (2.3)$$

We have accomplished the latter task in [18], getting the exponent $v = \frac{35}{48}$. Our arguments involve character sums in two variables

$$\sum_x \sum_y \left(\frac{x^2 - 1}{y} \right), \quad (2.4)$$

which we treat by Weil's and Burgess's estimates. It should be noticed that the character sum (2.4) does not live on any algebraic variety over a finite field to which one is tempted to use Deligne's theory. (The characteristic of the field depends on the variables!)

Since the Riemann hypothesis is true for the Selberg zeta-function $Z_T(s)$ (the complex zeros are $s_j = 1/2 + it_j$), one should expect the prime geodesic theorem to be true with the error term $O(x^{1/2+\varepsilon})$. Is there any conjecture about exponential sums which implies that? Here is one:

CONJECTURE 1. For $n, C, D \geq 1$ and $\varepsilon > 0$ we have $\sigma(n, C, D) \ll (nCD)^\varepsilon$ with the constant implied in \ll depending on ε alone.

Our conjecture constitutes a generalization of the Linnik-Selberg conjecture $\sigma(n, C, 0) \ll (nC)^\varepsilon$, which is sharp enough to imply the Ramanujan conjectures for both the finite and infinite places (if Γ is the congruence group $\Gamma_0(q)$, then the summation in (2.3) is restricted to $c \equiv 0 \pmod{q}$).

We close this section by expressing some thoughts to be supported in the following sections. Either the Selberg trace formula or the Kloosterman sums formula is a specialization of the spectral theorem. Since (2.1) is recognized as the limit case of (2.2) we have in a sense the following gradation: Spectral Theorem \succ Kloosterman Sums Formula \succ Selberg's Trace Formula. The Kloosterman sums formula holds a perfect balance between the amount of spectral information it carries and the potentiality for excavating it by means now available. These means are the methods of exponential sums in several variables. On using upper bounds for individual sums, one reaches the same limits as those set by formal Fourier analysis applied to the Selberg trace formula. A hypothetical variation in the sign of such exponential sums opens the way to pass over those limits. The possibility of doing so unconditionally is given by the free parameters m, n . A proper Fourier analysis in m, n decomposes the exponential sums and prepares them for estimation in each variable separately. This preliminary step takes advantage of an interplay between the finite and infinite places, which both contribute to the spectral side of the Kloosterman sums formula.

3. Small eigenvalues. The problem of estimating the first eigenvalue λ_1 occupies a central position in differential geometry. In the theory of automorphic functions it is particularly important to know whether there are eigenvalues $\lambda_j < \frac{1}{4}$. These eigenvalues will be called "exceptional" with the emphasis that they are not welcome. That the exceptional eigenvalues may exist was shown by B. Randol and was surely known to Selberg. In the case of compact, smooth quotients $\Gamma \backslash H$ of genus $g \geq 2$ we recall some interesting examples of R. Schoen, S. Wolpert, and S. T. Yau [33] having $\lambda_{2g-3} < \varepsilon$ and the general lower bound of

P. Buser [4] $\lambda_{4g-2} \geq \frac{1}{4}$. It is remarkable that the first eigenvalue λ_1 is bounded from above by an absolute constant $\lambda_1 \leq 2(g+1)/(g-1) \leq 6$. An extension to noncompact surfaces is given in [42].

In 1965 A. Selberg [36] stated the fundamental conjecture.

CONJECTURE 2. There are no exceptional eigenvalues for congruence groups.

The conjecture is known to be true for the modular group (H. Maass [27], W. Roelcke [31]), for triangle groups (P. Sarnak [32]), and for a few other groups having fundamental domains of relatively simple shape (M. Huxley [16]). The proofs are essentially geometrical. Selberg gave an equivalent formulation of his eigenvalue conjecture in terms of Kloosterman sums, since he showed that the zeta-function (2.3) admits analytic continuation up to $\operatorname{Re} s > \sigma$ ($\frac{1}{2} < \sigma < 1$) if and only if $\lambda_1 \geq \sigma(1 - \sigma)$; therefore Conjecture 1 implies Conjecture 2. From Weil's bound for Kloosterman sums,

$$|S(m, n; c)| \leq (m, n, c)^{1/2} c^{1/2} \tau(c), \quad (3.1)$$

Selberg got the remarkable bound $\lambda_1 \geq \frac{3}{16}$. Because of (3.1) Selberg's proof depends indirectly on the Riemann hypothesis for curves over finite fields. Perhaps it is interesting to mention that we [22] can prove the same bound $\lambda_1 \geq \frac{3}{16}$ without (3.1). What we use instead is an elementary inequality

$$\sum_{c < C} \sum_{q < Q} |S(n, n; qc)| \leq 4\tau_3(n)(C+Q)CQ(\log 4CQ)^2.$$

It is easy to find eigenvalues of $\Gamma_0(q)$ arbitrarily close to $\frac{1}{4}$ as q gets large; see [7]. In fact $\frac{1}{4}$ is an eigenvalue of Δ acting on $L^2(\Gamma_0(p) \backslash H, \chi_p)$ if p is prime $\equiv 1 \pmod{4}$ and the class number $h(p)$ of $Q(\sqrt{p})$ is greater than 1; see [26] and [38].

In practice one exceptional eigenvalue does not matter, but a large number of them may spoil the results. Therefore it is important to investigate the distribution of the exceptional points s_j in the segment $\frac{1}{2} < s_j < 1$ from a statistical point of view. A convenient way is to count all points s_j with certain weights such that the larger s_j is (more exceptional λ_j) the heavier the weight attached to s_j . The following statement, called the "density theorem," seems appropriate.

$$\sum_{1/2 < s_j < 1} |F|^{A(s_j - 1/2)} \ll_\varepsilon |F|^{1+\varepsilon}. \quad (3.2)$$

By methods very similar to those applied for the prime geodesic theorem we proved (3.2) with $A = \frac{12}{5}$; see [21]. Earlier we had, with J. Szmidt [19], $A = \frac{24}{11}$ and conditionally, assuming the Lindelöf hypothesis for Dirichlet's L -series, $A = 3$. The density theorem with $A = 2$ is quite easy to prove in two ways, either by using (3.1) [19] or by counting trivially the length spectrum in the Selberg trace formula for a proper test function [17]. Geometrical methods yield only $A = 0$ since the genus g of $\Gamma \backslash H$ is as large as the volume $|F|$.

4. Fourier coefficients of cusp forms as "characters." From now on we restrict our talk exclusively to the Hecke congruence groups $\Gamma = \Gamma_0(q)$. A few cusp forms were constructed explicitly by H. Maass [27], but by Weyl's law we

know that there are plenty of other forms. In fact any segment of $[\frac{1}{4}, \infty)$ contains $\gg q$ eigenvalues (counted with multiplicity) provided q is sufficiently large. Let us arrange the eigenvalues in nondecreasing order $\lambda_0 = 0 < \lambda_1 \leq \lambda_2 \leq \dots$ and choose an orthonormal basis $\{u_j\}_{j=1}^\infty$ of the space $L_0^2(\Gamma \backslash H)$ of cusp forms. Each of these functions u_j has the Fourier expansion (1.1). We normalize the Fourier coefficients $\rho_j(n)$ so that $\lambda_j(n) = \rho_j(n)\Gamma(s_j)|F|^{1/2}$ satisfy

$$\sum_{n \leq N} |\lambda_j(n)|^2 \sim \zeta(2)N \quad \text{as } N \rightarrow \infty$$

by the Rankin-Selberg convolution method. (Additionally we could require the forms u_j to be eigenfunctions of the Hecke operators T_n with $(n, q) = 1$ to gain a multiplicativity property for $\lambda_j(n)$, but it is not necessary.) Therefore one should think of $|\lambda_j(n)|$ as being bounded in average from below and above by absolute positive constants. This, of course, cannot be said about an individual number $\lambda_j(n)$; the upper bound $\lambda_j(n) \ll (\lambda_j n q)^\varepsilon$ is the content of the Ramanujan conjecture while a lower bound is unpredictable.

Even more unrealistic seems to be the problem of determining the sign of $\lambda_j(n)$. Certain information can be inferred from upper bounds for the linear forms

$$\mathcal{L}_j(\mathbf{a}) = \sum_{N < n \leq 2N} a_n \lambda_j(n),$$

where, the more general the coefficients $\mathbf{a} = (a_n)$ are permitted to be, the deeper our knowledge of λ_j will be. If we confront $\lambda_j(n)$ with the additive characters $a_n = e(\alpha n)$ we get an upper bound à la Hecke $\mathcal{L}_j(\mathbf{a}) \ll N^{1/2} \log N$ with the constant implied in \ll depending on u_j only, thus showing some orthogonality of the sequences $\lambda_j(n)$ and $e(\alpha n)$. For general linear forms we [5] proved the following mean-value theorem.

THEOREM (DESHOUILLEERS AND IWANIEC). *For any complex numbers a_n and any $\varepsilon > 0$ we have*

$$\sum_{|s_j| \leq T} |\mathcal{L}_j(\mathbf{a})|^2 \ll (N_\Gamma(T) + N^{1+\varepsilon}) \|\mathbf{a}\|^2 \quad (4.1)$$

with the constant implied in \ll depending on ε only.

Here we have $N_\Gamma(T) \ll q^{1+\varepsilon} T^2$. Similar statements are known in multiplicative number theory as large sieve inequalities. For example, for primitive Dirichlet characters we have Bombieri's inequality

$$\sum_{q \leq Q} \sum_{\chi \pmod{q}}^* \left| \sum_1^N a_n \chi(n) \right|^2 \ll (Q^2 + N) \|\mathbf{a}\|^2.$$

Our proof of (4.1) depends on estimates for bilinear forms in Kloosterman sums of the type

$$\sum_{N < m, n \leq 2N} a_m \bar{a}_n S(m, n; c) e\left(\frac{2\sqrt{mn}}{c}\right) \ll (cN)^{1/2+\varepsilon} \|\mathbf{a}\|^2$$

with $1 \leq c \leq N$. This inequality is stronger by a factor $N^{1/2}$ than the trivial one resulting from (3.1). It shows a great variation in the sign of $S(m, n, c)$ in terms of m and n .

If we restrict the summation in (4.1) to the exceptional points $\frac{1}{2} < s_j < 1$ we can do better by the introduction of weights in much the same fashion as we have done in the density theorem (3.2). We can also create an extra average over the level $q \leq Q$. The best we have got in this direction is the following weighted large sieve inequality.

THEOREM (DESHOILLERS AND IWANIEC). *For any complex numbers a_n , $X \geq 1$, and $\varepsilon > 0$ we have*

$$\frac{1}{Q} \sum_{q \leq Q} \sum_{1/2 < s_j < 1} X^{4(s_j - 1/2)} |\mathcal{L}_j(\mathbf{a})|^2 \ll (NQ)^\varepsilon (Q + NX) \|\mathbf{a}\|^2. \quad (4.2)$$

Still a stronger inequality can be proved for special linear forms $\mathcal{L}_j(\mathbf{a})$ whose coefficients a_n are additive characters.

The fundamental tools of analytic number theory are characters, like the additive ones $e(\alpha n)$ and the Dirichlet characters $\chi \pmod{q}$. Recently the meaning of characters has been loosely extended to additive or multiplicative functions which possess a sort of orthogonality property; for example $f(n) = n^{it}$, $t \in \mathbf{R}$, are regarded as characters in the theory of Dirichlet's polynomials. In accordance with this convention we give the Fourier coefficients of cusp forms the status of characters.

The orthogonality of characters is exploited to select those elements of a sequence under an investigation which possess certain properties. For instance L. Dirichlet used his multiplicative characters $\chi \pmod{q}$ to pick up primes in an arithmetic progression. The new characters $\lambda_j(n)$ have the potential to solve a diophantine equation of the type $ad - bc = n$. Such an equation with various constraints on the variables a, b, c, d occurs in numerous problems in number theory. The constraints can be quite serious, especially when the initial problem involves prime numbers. Usually we transform the constraints into congruence conditions by using sophisticated methods of analytic number theory and then we relate the equation $ad - bc = n$ to the spectral theory of $\Gamma_0(q)$. For example, in the proof of (0.1) the constraints are released by dispersion methods.

5. Modular forms of half-integral weight. In this section we consider a holomorphic modular form f for $\Gamma_0(q)$ of weight $k = \frac{1}{2} + l \geq \frac{3}{2}$ and with a multiplier system given by the theta series. Such forms occur in studies of quadratic forms and elliptic curves. Two of the most fundamental results about metaplectic forms are the Shimura correspondence [37] and the Waldspurger theorem [41]. Our interest is to estimate the Fourier coefficients. A proper analogue of the Ramanujan conjecture asserts that for any $\varepsilon > 0$

$$a_n \ll n^{(k-1)/2+\varepsilon} \quad (5.1)$$

provided n is square-free (otherwise the statement would be false). Let us recall that Deligne's results refer to holomorphic forms of integral weight. The bound

(5.1) with $\varepsilon > \frac{1}{4}$ is easy to obtain either by using Weil's estimate for the relevant Kloosterman sums or directly from the Waldspurger theorem

$$|a_n|^2 = cn^{k-1} L_g(k - \tfrac{1}{2}, \chi_n), \quad (5.2)$$

where g is the image of f under the Shimura correspondence (so g is a holomorphic form of integral weight $2k - 1$) and L_g is the L -series of g twisted by real character $\chi_n \pmod{n}$ evaluated at the center of the critical strip $s = k - \frac{1}{2}$. The functional equation and a convexity argument yield $L_g(k - \frac{1}{2}, \chi_n) \ll n^{1/2} \log^2 n$ whence (5.1) with $\varepsilon > \frac{1}{4}$. One should observe that the Ramanujan conjecture is equivalent to the Lindelöf hypothesis for $L_g(k - \frac{1}{2}, \chi_n)$.

It is interesting from the point of view of analytic number theory to extend this discussion to the nonholomorphic Maass forms and to the nonholomorphic Eisenstein series. In the latter case the twisted L -series is essentially a product of two Dirichlet's L -series with real character χ_n and with both arguments on the critical line $\operatorname{Re} s = \frac{1}{2}$; see [12]. Thus the Ramanujan conjecture for the Fourier coefficients of Eisenstein series is nothing other than the Lindelöf hypothesis for Dirichlet's L -series in the n -aspect. Though the Ramanujan conjectures are far out of reach, we remark that in the case of integral weight there is at least a plausible program of Langlands to attack the conjecture by lifting to modular forms for groups of high rank. But in the case of half-integral weight one cannot even start with the program because the Hecke operators annihilate a_n .

Some techniques of exponential sums seem to be promising. Let us consider the Petersson formula for the n th Fourier coefficient $\hat{P}_n(n, k)$ of the Poincaré series $P_n(z, k)$

$$\hat{P}_n(n, k) = 1 + 2\pi i^{-k} \sum_{c \equiv 0 \pmod{q}} c^{-1} S(n, n; c) J_{k-1} \left(\frac{4\pi n}{c} \right).$$

Since $P_n(z, k)$ span the space of cusp forms we also have $|a_n|^2 \ll \hat{P}_n(n, k)$. Here, the very special structure of Kloosterman sums $S(n, n; c)$ intrinsic to modular forms of half-integral weight allows us to evaluate them explicitly and then reduces the problem to bounding bilinear forms of the type

$$\mathfrak{B}(\mathbf{X}, \mathbf{Y}) = \sum_{A < a \leq 2A} \sum_{B < b \leq 2B} x_a y_b \cos \left(2\pi n \left(\frac{\bar{a}}{b} - \frac{\bar{b}}{a} \right) \right),$$

where \bar{a} , \bar{b} stand for solutions to $a\bar{a} \equiv 1 \pmod{b}$ and $b\bar{b} \equiv 1 \pmod{a}$ respectively. We know how to deal with $\mathfrak{B}(\mathbf{X}, \mathbf{Y})$ when A and B/A are both large. In order to obtain a nontrivial and unconditional result we must avoid the case $A \cong B$. To this end we apply a simple, but cute, idea of embedding the space of cusp forms for $\Gamma_0(q)$ into the space of cusp forms for a subgroup $\Gamma_0(qr)$ of large level qr . Then by averaging over r we gain some flexibility in the structure of the bilinear forms $\mathfrak{B}(\mathbf{X}, \mathbf{Y})$ that permits us to rearrange the variables so to avoid the case $A \cong B$. In this way we [23] obtained (5.1) with any $\varepsilon > \frac{3}{14}$. The improvement over the trivial result $\varepsilon > \frac{1}{4}$ is small indeed; nevertheless it is vital for problems

such as the equidistribution of integral points on the sphere $x^2 + y^2 + z^2 = n$ and on allied surfaces.

6. Epilogue. The very first symptoms of the “Kloostermania” were observed after Kuznetsov’s striking result

$$\sum_{c \leq C} c^{-1} S(m, n; c) \ll_{\varepsilon, m, n} C^{1/6+\varepsilon}. \quad (6.1)$$

This was proved by a trivial application of his Kloosterman sums formula [24] (see also the later publications [5, 11]). Our results about the spectrum of Δ and about the Fourier coefficients of cusp forms enforce (6.1) substantially; see [5]. The most valuable is the weighted large sieve inequality (4.2) without which many estimates would be useless unless one assumes the Selberg Conjecture 2. While Selberg’s conjecture seems to be out of reach, the statistical theorems deserve further development. This is a modest but realistic program. Deshouillers and the speaker have found many practical situations in which the statistical theorems produce the same results as would the Selberg eigenvalue conjecture. If you cannot move mountains, walk around them!

REFERENCES

1. E. Bombieri, *On the large sieve*, *Mathematika* **12** (1965), 201–225.
2. E. Bombieri, J. B. Friedlander, and H. Iwaniec, *Primes in arithmetic progressions to large moduli*. II, *Math. Ann.* (to appear).
3. R. W. Burgess, *Fourier coefficients of cusp forms*, *Invent. Math.* **45** (1978), 1–18.
4. P. Buser, *Riemannsche Flächen mit Eigenwerten in $(0, \frac{1}{4})$* , *Comment. Math. Helv.* **52** (1977), 25–34.
5. J.-M. Deshouillers and H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms*, *Invent. Math.* **70** (1982), 219–288.
6. —, *Power mean-values for the Riemann zeta-function*, *Mathematika* **29** (1982), 202–212.
7. —, *Accumulation of the eigenvalues of the hyperbolic Laplacian at $\frac{1}{4}$* , Preprint, 1983.
8. —, *The non-vanishing of the Rankin-Selberg zeta-functions at special points*, *Selberg Trace Formula and Related Topics*, *Contemp. Math.*, vol. 53, Amer. Math. Soc., Providence, R. I., 1986, pp. 51–95.
9. J.-M. Deshouillers, H. Iwaniec, R. S. Phillips, and P. Sarnak, *Maass cusp forms*, *Proc. Nat. Acad. Sci. U.S.A.* **82** (1985), 3533–3534.
10. E. Fouvry, *Théorème de Brun-Titchmarsh; application au théorème de Fermat*, *Invent. Math.* **79** (1985), 383–407.
11. D. Goldfeld and P. Sarnak, *Sums of Kloosterman sums*, *Invent. Math.* **71** (1983), 243–250.
12. D. Goldfeld and J. Hoffstein, *Eisenstein series of $\frac{1}{2}$ -integral weight and the mean value of real Dirichlet L-series*, *Invent. Math.* **80** (1985), 185–208.
13. R. Gupta and R. Murty, *A remark on Artin’s conjecture*, *Invent. Math.* **78** (1984), 127–130.
14. D. R. Heath-Brown, *Artin’s conjecture for primitive roots*, *Quart. J. Math.* (2) **37** (1986), 27–38.
15. D. A. Hejhal, *The Selberg trace formula*. Vol. 2, *Lecture Notes in Math.*, vol. 1001, Springer-Verlag, 1983.
16. M. Huxley, *Introduction to Kloostermania*, *Elementary and Analytic Theory of Numbers*, Banach Center Publ., vol. 17, PWN, Warsaw, 1985.

17. —, *Exceptional eigenvalues and congruence subgroups*, Selberg Trace Formula and Related Topics, Contemp. Math., vol. 53, Amer. Math. Soc., Providence, R. I., 1986, pp. 341–349.
18. H. Iwaniec, *Prime geodesic theorem*, J. Reine Angew. Math. **349** (1984), 136–159.
19. H. Iwaniec and J. Szmíd, *Density theorems for exceptional eigenvalues of Laplacian for congruence groups*, Elementary and Analytic Theory of Numbers, Banach Center Publ., vol. 17, PWN, Warsaw, 1985.
20. H. Iwaniec and J. Pintz, *Primes in short intervals*, Mh. Math. **98** (1984), 115–143.
21. H. Iwaniec, *Character sums and small eigenvalues for $\Gamma_0(p)$* , Glasgow Math. J. **27** (1985), 99–116.
22. —, *Selberg's lower bound for the first eigenvalue for $\Gamma_0(N)$* , Preprint, 1986.
23. —, *Fourier coefficients of modular forms of half-integral weight*, Invent. Math. (to appear).
24. N. V. Kuznetsov, *The Petersson conjecture for cusp forms of weight zero and the Linnik conjecture*, Preprint, Khabarovsk, 1977. (Russian)
25. —, *The arithmetic form of Selberg's trace formula and the distribution of the norms of the primitive hyperbolic classes of the modular group*, Preprint, Khabarovsk, 1978. (Russian)
26. H. Maass, *Über eine neue Art von nichtanalytischen automorphen Funktionen und die Bestimmung Dirichletscher Reichen durch Funktionalgleichungen*, Math. Ann. **121** (1949), 141–183.
27. —, *Lectures on modular functions of one complex variable*, Tata Inst. Lecture Notes, no. 29, Tata Inst. Func. Res., Bombay, 1964.
28. R. S. Phillips and P. Sarnak, *The Weyl theorem and the deformation of discrete groups*, Comm. Pure Appl. Math. **38** (1985), 853–866.
29. —, *On cusp forms for cofinite subgroups of $\mathrm{PSL}(2, \mathbb{R})$* , Invent. Math. **80** (1985), 339–364.
30. P. Sarnak, *On cusp forms*, Selberg Trace Formula and Related Topics, Contemp. Math., vol. 53, Amer. Math. Soc., Providence, R. I., 1986, pp. 393–407.
31. W. Roelcke, *Über die Wellengleichung bei Grenzkreisgneppen erster Art*, S.-B. Heidelberger Akad. Wiss Math. Nat. Kl., 1956, 4. Abh.
32. P. Sarnak, *Prime geodesic theorems*, Ph.D. Thesis, Stanford University, 1980.
33. R. Schoen, S. Wolpert, and S. T. Yau, *Geometric bounds on the low eigenvalues of a compact surface*, Proc. Sympos. Pure Math., vol. 36, Amer. Math. Soc., Providence, R. I., 1980, pp. 279–285.
34. A. Selberg, *Göttingen lectures*, 1954 (unpublished).
35. —, *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series*, J. Indian Math. Soc. (N.S.) **20** (1956), 47–87.
36. —, *On the estimation of Fourier coefficients of modular forms*, Proc. Sympos. Pure Math., vol. 8, Amer. Math. Soc., Providence, R. I., 1965, pp. 1–15.
37. G. Shimura, *On modular forms of half-integral weight*, Ann. of Math. (2) **97** (1973), 440–481.
38. L. A. Takhtajan and A. I. Vinogradov, *The Gauss-Hasse hypothesis on real quadratic fields with class number one*, J. Reine Angew. Math. **335** (1982), 40–86.
39. A. B. Venkov, *Spectral theory of automorphic functions*, Trudy Mat. Inst. Steklov. **153** (1981), 1–170. (Russian)
40. A. I. Vinogradov, *On the density hypothesis for Dirichlet L -functions*, Izv. Akad. Nauk SSSR **29** (1965), 903–934. (Russian)
41. J.-L. Waldspurger, *Correspondences de Shimura et Shintani*, J. Math. Pures Appl. **59** (1980), 1–133.
42. P. Zograf, *Fuchsian groups and small eigenvalues of the Laplace operator*, Zap. Nauchn. Sem. Leningrad. Otdel. Inst. Steklov (LOMI) **122** (1982), 24–29. (Russian)

Unimodular Lattices and Self-dual Codes

HELMUT V. KOCH

The main goal of my talk is to show how to use weighted theta functions in the theory of unimodular lattices and self-dual codes. Though weighted theta functions were studied at the end of the thirties by Schoeneberg [10] and Hecke [2], their usefulness for the theory of unimodular lattices was demonstrated only in 1978 by B. B. Venkov [12].

1. Notations and definitions. For any field K we denote by $\langle x, y \rangle$ the standard scalar product in K^n , i.e.,

$$\langle x, y \rangle := x_1 y_1 + \cdots + x_n y_n$$

for $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ in K^n . The elements of \mathbf{F}_2^n , considered as functions on $\{1, \dots, n\}$, will be identified with their support; hence \mathbf{F}_2^n will be considered as the system of all subsets of $\{1, \dots, n\}$. We put $\mathbf{1} := \{1, \dots, n\}$, $\mathbf{0} := \emptyset$. The cardinality $w(x) = |x|$ of $x \in \mathbf{F}_2^n$ is called the *weight* of x .

A lattice G in \mathbf{Q}^n is called *even* if $\langle x, x \rangle \in 2\mathbf{Z}$ for all $x \in G$. This implies $\langle x, y \rangle \in \mathbf{Z}$ for all $x, y \in G$. The lattice G is called *unimodular* if it is equal to its *dual lattice* $G^0 := \{x \in \mathbf{Q}^n \mid \langle x, y \rangle \in \mathbf{Z} \text{ for all } y \in G\}$. Even unimodular lattices exist only if $8|n$.

Let C be a binary linear code of *length* n : $C \subset \mathbf{F}_2^n$. Then C is called *doubly-even* if $4|w(x)$ for all $x \in C$. The code C is called *self-dual* if it is equal to its *dual code* $C^\perp := \{x \in \mathbf{F}_2^n \mid \langle x, c \rangle = 0 \text{ for all } c \in C\}$. Doubly-even self-dual codes exist only if $8|n$ and contain always $\mathbf{1}$. The symmetric group S_n operates naturally on \mathbf{F}_2^n . Two codes C and C' , in \mathbf{F}_2^n are called *equivalent* if there is a $\Pi \in S_n$ with $C' = \Pi C$. The automorphism group $\text{Aut } C$ of C is defined by $\text{Aut } C := \{\Pi \in S_n \mid \Pi C = C\}$.

2. The connection between even lattices in \mathbf{Q}^n and doubly-even codes in \mathbf{F}_2^n . Usually one speaks about analogies between properties of even lattices and doubly-even codes; in this section of my talk I am going to show that actually the theory of doubly-even codes is a subtheory of the theory of even lattices.

Let G be an even lattice in \mathbf{Q}^n . We put $G_h := \{x \in G \mid \langle x, x \rangle = h\}$. G_2 is a root system with irreducible components of the form A_m , D_m , E_6 , E_7 , E_8 . We call G_2 complete if it generates the vector space \mathbf{Q}^n . Then $Q(G_2) := (G_2)_{\mathbf{Z}}^0 / (G_2)_{\mathbf{Z}}$ is a finite abelian group. Let $\Gamma(G_2)$ be the subgroup of $\text{Aut}(Q(G_2))$ induced by the isometry group of $(G_2)_{\mathbf{Z}}^0$ and $l_{G_2}: Q(G_2) \rightarrow \mathbf{Q}$ the *length function* defined by

$$l_{G_2}(\xi) := \min\{\langle x, x \rangle \mid x \in \xi\} \quad \text{for } \xi \in (G_2)_{\mathbf{Z}}^0 / (G_2)_{\mathbf{Z}}.$$

PROPOSITION 1 [12]. *Let G_2 be a root system of rank n with irreducible components of the form A_m , D_m , E_6 , E_7 , E_8 . There is a one-to-one correspondence between equivalence classes of even lattices in \mathbf{Q}^n with root system G_2 and $\Gamma(G_2)$ -orbits of subgroups A in $Q(G_2)$ with $l_{G_2}(\alpha) \in 2\mathbf{Z} - \{2\}$ for $\alpha \in A$.*

Let C be a doubly-even code of length n : $C \subset \mathbf{F}_2^n$. The set $C_4 := \{c \in C \mid w(c) = 4\}$ is called the *tetrad system* of C . A tetrad system can be decomposed into irreducible systems, which up to equivalence have the form

$$d_{2k} := (\{1, 2, 3, 4\}, \{1, 2, 5, 6\}, \dots, \{1, 2, 2k-1, 2k\})_4$$

for $k = 2, 3, \dots$,

$$e_7 := (\{1, 2, 3, 4\}, \{1, 2, 5, 6\}, \{1, 3, 5, 7\})_4,$$

$$e_8 := (\{1, 2, 3, 4\}, \{1, 2, 5, 6\}, \{1, 2, 7, 8\}, \{1, 3, 5, 7\})_4.$$

Analogously to the case of lattices we define the abelian group $Q(C_4) := (C_4)^\perp / (C_4)$, the subgroup $\Gamma(C_4)$ of $\text{Aut}(Q(C_4))$ induced by $\text{Aut}((C_4)^\perp) \subset S_n$, and the *weight function* $w_{C_4}: Q(C_4) \rightarrow \mathbf{Z}$ defined by

$$w_{C_4}(\eta) := \min\{w(y) \mid y \in \eta\} \quad \text{for } \eta \in (C_4)^\perp / (C_4).$$

PROPOSITION 2. *Let C_4 be a tetrad set in \mathbf{F}_2^n . There is a one-to-one correspondence between equivalence classes of doubly-even codes in \mathbf{F}_2^n with tetrad system C_4 and $\Gamma(C_4)$ -orbits of subgroups A of $Q(C_4)$ with $w_{C_4}(\alpha) \in 4\mathbf{Z} - \{4\}$ for $\alpha \in A$.*

Now we associate to each doubly-even code C in \mathbf{F}_2^n an even lattice in \mathbf{Q}^n : Let $H \subset \mathbf{Q}^n$ be a fixed lattice with orthogonal basis h_1, \dots, h_n such that $\langle h_i, h_i \rangle = 2$ for $i = 1, 2, \dots, n$. In other words, H is a lattice of the form (nA_1) . Then let $\phi(C)$ be the lattice generated by H and by the vectors $\frac{1}{2} \sum_{i \in c} h_i$ for all $c \in C$. Then $\phi(C)$ is an even lattice in \mathbf{Q}^n .

THEOREM 3. *ϕ induces a one-to-one correspondence between equivalence classes of doubly-even codes C in \mathbf{F}_2^n and equivalence classes of even lattices in \mathbf{Q}^n with root systems containing nA_1 . Self-dual codes correspond to unimodular lattices.*

PROOF. This follows from Propositions 1 and 2 together with the comparison of root systems and tetrad systems. A lot of properties have to be checked, but every check has the character of an exercise.

3. The basic equations for even unimodular lattices. Let $P_\nu(x, y)$ be the zonal spherical function of degree ν in \mathbf{R}^n and let G be an even unimodular lattice in \mathbf{Q}^n . Then the *weighted theta functions*

$$\theta_G(z, \nu, y) := \sum_{x \in G} P_\nu(x, y) e^{\pi i z \langle x, x \rangle}, \quad \nu = 2, 4, 6, \dots, y \in \mathbf{R}^n,$$

are parabolic modular forms of weight $n/2 + \nu$ for the full modular group [2, 10].

If $n = 24$ and $\nu = 2$, then the weight is 14 but there is no nontrivial parabolic modular form of this weight. Hence

$$\sum_{x \in G_h} P_2(x, y) = \sum_{x \in G_h} \langle x, y \rangle^2 - |G_h| \frac{h \langle y, y \rangle}{24} = 0 \quad \text{for } h = 2, 4, \dots, y \in \mathbf{R}^{24}, \quad (1)$$

which is a very powerful relation as we shall soon see.

If $n = 32$, we get four equations relating the elements of G_4 with the elements of G_2 . (The spaces of parabolic modular forms of weight 18, 20, 22, 26 have dimension 1.) We call these equations the basic equations [13].

4. The basic equations for extremal codes. Now let C be an *extremal code* of length n , i.e., a doubly-even self-dual code in \mathbf{F}_2^n of minimal weight $4\lfloor n/24 \rfloor + 4$. By the theorem of MacWilliams-Gleason these are the doubly-even self-dual codes in \mathbf{F}_2^n with highest possible minimal weight, hence the most interesting. Their *weight enumerator*

$$f_C(x, y) := \sum_{c \in C} x^{w(c)} y^{n-w(c)}$$

is uniquely determined by n . For the first n one finds the following values for the number of code words of minimal weight $h = 4\lfloor n/24 \rfloor + 4$:

n	8	16	24	32	40	48
h	4	4	8	8	8	12
$ C_h $	14	28	759	620	285	17296
$ C_h $	2·7	2 ² ·7	3·11·23	2 ² ·5·31	3·5·19	2 ⁴ ·23·47

In fact Mallows and Sloane [6] showed

$$\begin{aligned} |C_h| &= \binom{5r-2}{r-1} \binom{n}{5} \binom{h}{5}^{-1} & \text{if } n = 24r, \\ |C_h| &= \frac{1}{4} n(n-1)(n-2)(n-4) \frac{(5r)!}{r!h!} & \text{if } n = 24r + 8, \\ |C_h| &= \frac{3}{2} n(n-2) \frac{(5r+2)!}{r!h!} & \text{if } n = 24r + 16, \end{aligned}$$

and Mallows, Odlyzko, and Sloane [5] proved that extremal codes do not exist if n is sufficiently high. The smallest n for which one does not know at present whether an extremal code of length n exists is 72.

We put $n = 24r + 8s$, $s = 0, 1, 2$. For given $a \in \mathbb{F}_2^n$ let

$$\theta_{\phi(C)} \left(z, \nu, \sum_{i \in a} h_i \right) = \alpha_1 q + \cdots + \alpha_r q^r + \cdots \quad \text{with } q := e^{2\pi i z}$$

be the weighted theta function of $\phi(C)$.

Then, for $\nu = 2, \dots, 10 - 4s, 14 - 4s$, $\theta_{\phi(C)}$ lies in a space of parabolic modular forms of dimension $\leq r$. This implies that the coefficients α_i are linear combinations of $\alpha_1, \dots, \alpha_r$ with complex coefficients. But α_i , $i = 1, \dots, r$, depends only on a and is in fact a polynomial in $w(a)$ of degree $\leq \nu/2$. Taking into account the well-known structure of the zonal spherical functions $P_\nu(x, y)$ we get the following result:

THEOREM 4 [4]. *Let C be an extremal code of length $n = 24r + 8s$, $s = 0, 1, 2$. Let C_h be the set of words in C with minimal weight $h = 4r + 4$ and let a be an arbitrary element of \mathbb{F}_2^n . Then*

$$\sum_{x \in C_h} w(x \cap a)^i = f_i(w(a)) \quad \text{for } i = 1, 2, \dots, t := 5 - 2s, \quad (2)$$

and

$$\sum_{x \in C_h} (w(x \cap a)^i - g_i(w(a))w(x \cap a)^{i-1}) = f_i(w(a)) - g_i(w(a))f_{i-1}(w(a)) \quad (3)$$

for $i = 7 - 2s$ with

$$g_7(j) := 14 + \frac{7}{6}j, \quad g_5(j) := \frac{20}{3} + \frac{5}{6}j, \quad g_3(j) := 2 + \frac{1}{2}j.$$

$f_i(\xi)$ is a polynomial of degree i of the form

$$f_i(\xi) := |C_h| \sum_{\kappa=1}^i \binom{h}{\kappa} \binom{n}{\kappa}^{-1} S_{i\kappa} \xi(\xi-1) \cdots (\xi-\kappa+1),$$

where $S_{i\kappa}$ is defined by

$$S_{i0} = 0 \quad \text{for } i > 0, \quad S_{ii} = 1 \quad \text{for } i \geq 0,$$

$$S_{i+1, \kappa} = \kappa S_{i\kappa} + S_{i, \kappa-1} \quad \text{for } i > \kappa > 0.$$

The numbers $S_{i\kappa}$ are called Stirling numbers of the second kind.

We call the $t+1$ equations (2), (3) the *basic equations* of the extremal code C . Let

$$u_j(a) := |\{x \in C_h \mid w(a \cap x) = j\}|.$$

The basic equations are equivalent to the following system of equations for the $u_j(a)$, which is more convenient for computations:

$$\sum_{j=\nu}^{w(a)} \binom{j}{\nu} u_j(a) = |C_h| \binom{h}{\nu} \binom{n}{\nu}^{-1} \binom{w(a)}{\nu}, \quad \nu = 1, \dots, t, \quad (4)$$

$$\sum_{j=t+1}^{w(a)} \left(\binom{j}{t+1} (w(a) - j) + (t-4) \binom{j}{t+2} \right) u_j(a) = \binom{w(a)}{t+2} \kappa_t \quad (5)$$

with

$$\begin{aligned} \kappa_5 &:= \binom{5r-2}{r-1} = |C_h| \binom{h}{5} \binom{n}{5}^{-1}, \\ \kappa_3 &:= \binom{5r}{r}, \quad \kappa_1 := 3 \binom{5r+2}{r}. \end{aligned}$$

We remember that $n = 24r + 8s$, $t = 5 - 2s$, $s = 0, 1, 2$.

(4) implies the

THEOREM OF ASSMUS-MATTSON. *Let $w(a) = t$. Then $u_t(a)$ is independent of the choice of a , i.e., C_h is a t -block design.*¹

(5) implies the

THEOREM OF VENKOV [15]. *Let $w(a) = t + 2$. Then $u_{t+1}(a) + (t-4)u_{t+2}(a) = \kappa_t$ is independent of the choice of a .*

On the other hand one deduces purely combinatorically (4) and (5) resp., from the theorem of Assmus-Mattson and from the theorem of Venkov resp.

5. Classification of even unimodular lattices and doubly-even self-dual codes. Even unimodular lattices can be classified up to dimension 24. The number of inequivalent even unimodular lattices in dimension 32 is greater than 80,000,000 [11]. Doubly-even self-dual codes are classified up to length 32. The number of inequivalent doubly-even self-dual codes of length 40 is greater than 17,000.

The theory of even unimodular lattices in euclidean n -space began with investigations of Korkin and Zolotarev in 1873 and Minkowski in 1882, who discovered the even unimodular lattice (E_8) being generated by the root system E_8 . The lattice (E_8) corresponds to the $(8, 4)$ -Hamming code (e_8) , i.e., $(E_8) = \phi((e_8))$. Mordell [7] showed in 1938 that up to isometry (E_8) is the unique even unimodular lattice of dimension 8. In 1941 Witt [16] classified even unimodular lattices of dimension 16. Up to isometry there are two such lattices with root system $2E_8$, D_{16} , respectively. The even unimodular lattices of dimension 24 were classified by Niemeier [8] in 1973, who among other things used methods developed by M. Kneser. Niemeier's somewhat enigmatic list of 24 lattices was explained by B. B. Venkov [12], who also gave a shorter proof by means of weighted theta functions (see §6). Doubly-even self-dual codes of length 32 were classified by Conway and Pless [1] in 1980 (see §7).

As an example we consider *extended quadratic residue codes*: Let p be an odd prime. The extended quadratic residue code $C(p)$ is a linear subspace in \mathbb{F}_2^{p+1} .

¹The theorem of Assmus-Mattson can be proved for arbitrary weight h by means of the weighted theta functions of $\phi(C)$.

The places of the code are the points of the projective line $\mathbf{P}_1(\mathbf{F}_p) = \mathbf{F}_p \cup \{\infty\}$, and a basis of $C(p)$ is given by the words

$$w_\nu := \{\nu + i^2 \mid i \in \mathbf{F}_p^x\} \cup \{\infty\}, \quad \nu = 0, \dots, (p-1)/2.$$

$C(p)$ is a doubly-even self-dual code for all primes $p \equiv -1 \pmod{8}$, and $C(p)$ is an extremal code for $p = 7, 23, 31, 47$. The automorphism group of $C(p)$ contains $\text{PSL}_2(\mathbf{F}_2)$, and one conjectures that $\text{Aut}(C(p)) = \text{PSL}_2(\mathbf{F}_p)$ for $p \geq 31$ [9].

Now we look more closely at the lattices in dimension 24 and at extremal codes of length 32.

6. Even unimodular lattices of dimension 24 [12]. Let G be an even unimodular lattice of dimension 24. I have already mentioned that Venkov [12] with the method of weighted modular forms found the equation

$$\sum_{x \in G_2} \langle x, y \rangle^2 = |G_2| \frac{2\langle y, y \rangle}{24} \quad \text{for } y \in \mathbf{R}^{24}.$$

This equation easily implies the following facts:

- (a) The root system G_2 of the lattice G is empty or it generates \mathbf{Q}^{24} .
- (b) All irreducible components X of G_2 have the same Coxeter number $c(X)$.
 $(c(A_m) = m + 1, c(D_m) = 2m - 2, c(E_6) = 12, c(E_7) = 18, c(E_8) = 30.)$

It is an easy exercise to deduce from (a) and (b) that there are only the following 24 possibilities for root systems of even unimodular lattices G of dimension 24: $G_2 = \emptyset, 24A_1, 12A_2, 8A_3, 6A_4, 4A_6, 3A_8, 2A_{12}, A_{24}, 6D_4, 4D_6, 3D_8, 2D_{12}, D_{24}, 4E_6, 3E_8, 4A_5 + D_4, 2A_7 + 2D_5, 2A_9 + D_6, A_{15} + D_9, 2E_7 + D_{10}, E_7 + A_{17}, E_8 + D_{16}, E_6 + D_7 + A_{11}$. It remains to prove that each of these root systems belongs exactly to one lattice. This has essentially to be done in the same way as in Niemeier's work [8] with some simplifications by means of coding theory.

According to Theorem 3 the following 9 root systems correspond to doubly-even self-dual codes of length 24: $24A_1, 6D_4, 4D_6, 3D_8, 2D_{12}, D_{24}, 3E_8, 2E_7 + D_{10}, E_8 + D_{16}$.

In particular $24A_1$ corresponds to the unique code without tetrads, i.e., the extremal code, which is the Golay code $C(23)$.

7. Doubly-even self-dual codes of length 32. Conway and Pless [1] classified the doubly-even self-dual codes of length 32 by means of the mass formula

$$\sum_C \frac{1}{|\text{Aut } C|} = \frac{1}{n!} \prod_{j=0}^{(n-4)/2} (2^j + 1).$$

One has to find, for instance "by divination," all the 85 codes, and one has to compute their automorphism groups. Then the mass formula shows that one found all codes. This method is very laborious and actually [1] is only a description of the procedure. I wanted to give a shorter proof by means of the method of weighted theta functions, and I considered the most interesting case of extremal codes [4]. In fact I only needed the theorem of Assmus and Mattson and the following *balance principle*, the proof of which is elementary and easy.

BALANCE PRINCIPLE. Let $C \subset \mathbb{F}_2^n$ be a self-dual code of length n and let $a \in \mathbb{F}_2^n$ be arbitrary. Then

$$\frac{w(a)}{2} - \dim C(a) = \frac{w(\mathbf{1} + a)}{2} - \dim C(\mathbf{1} + a),$$

where $\mathbf{1} := \{1, \dots, n\}$ and

$$C(a) := \{x \in C \mid x \subset a\} \subset \mathbb{F}_2^a.$$

The balance principle follows from the exact sequence

$$\{0\} \rightarrow C(\mathbf{1} + a) \rightarrow C \xrightarrow{\text{Pr}_a} C(a)^\perp \rightarrow \{0\}, \quad (6)$$

where Pr_a is the projection on a . (6) together with the dual sequence implies a canonical isomorphism

$$\psi: C(a)^\perp / C(a) \rightarrow C(\mathbf{1} + a)^\perp / C(\mathbf{1} + a). \quad (7)$$

On the other hand, C is determined by $C(a)$, $C(\mathbf{1} + a)$, and ψ .

The main idea of my proof that there are exactly five extremal codes of length 32 consists in the consideration of the *configurations* of a code:

If C is an extremal code of length $n = 24r + (5 - t)4$, $t = 5, 3, 1$, then by the theorem of Assmus and Mattson the set C_h of words in C of minimal weight h form a t -block design. For any $a \in \mathbb{F}_2^n$ with $w(a) = t$ we get a (nonlinear) subcode $C^a := \{x \in C_h \mid a \subset x\}$. We call C^a a *configuration of the code*.

For $n = 32$ one has $|C^a| = 7$. It is easy to see that up to equivalence there are 7 codes in \mathbb{F}_2^{32} which are candidates for configurations of extremal codes. It is also rather easy to classify extremal codes of length 32 by means of these candidates for configurations.

8. Even unimodular lattices in \mathbb{Q}^{32} . I have already mentioned that there are more than 80 million even unimodular lattices in \mathbb{Q}^{32} . Hence it is not possible to classify them. Venkov studied these lattices in his papers [13, 14] and got interesting results about them:

One may ask whether such a lattice G is generated by its vectors in $G_2 \cup G_4$. Venkov showed by means of the basic equations that this is always the case if $G_2 = \emptyset$, i.e., G has no roots. More generally, there are only a few cases of lattices which are not generated by their vectors in $G_2 \cup G_4$, and these cases can be described explicitly. The situation is similar to the classification of algebraic surfaces, where we have some special classes of surfaces which can be described in detail, and we have the surfaces of general type which have a characteristic property but we know little about them beside this.

9. Extremal codes of length 48. At the end of my talk I want to say something about extremal codes of length 48. Such codes are of special interest from the combinatorial point of view since they lead to 5-block designs by the theorem of Assmus and Mattson. So far one knows only one extremal code of length 48, the extended quadratic residue code $C(47)$. Huffman [3] proved that

an extremal code of length 48 with an automorphism of odd order must be the extended quadratic residue code. I would like to classify extremal codes of length 48 using among other tools the theorem of Venkov, and I am going to explain my method.

We take a code word a with $w(a) = 24$. Then it is easy to see that there are four possibilities for the dimension of $C(a)$:

$$\dim C(a) = \dim C(1 + a) \in \{1, 2, 3, 4\}.$$

Moreover the structure of the code $C(a)$ is uniquely determined by its dimension. We put $g(C) := \max\{\dim C(a) \mid a \in C_{24}\}$.

It is easy to see that $g(C) \geq 2$. But it took me much time to show (by hand) that $g(C) > 2$.

An extremal code C with $g(C) = 2$ would be very interesting because it would have the following property:

For any $a \in \mathbb{F}_2^{48}$ with $w(a) = 6$, $|C^a| := |\{x \in C_{12} \mid a \subset x\}| \in \{0, 2\}$.

We want to show that this is impossible. We take two words $x_1, x_2 \in C_{12}$ with $w(x_1 \cap x_2) = 6$. Without loss of generality let $x_1 = \{1, \dots, 12\}$, $x_2 = \{1, \dots, 6, 13, \dots, 18\}$. Then there is exactly one word $a_{ij} \in C_{12}$ such that

$$(\{1, \dots, 6\} - \{i\}) \cup \{j\} \in a_{ij} \quad \text{and} \quad a_{ij} \neq x_1, \quad i = 1, \dots, 6, \quad j = 7, \dots, 12.$$

Let $b_{ij} := a_{ij} - ((\{1, \dots, 6\} - \{i\}) \cup \{j\})$. Each b_{ij} consists of 6 numbers α with $12 < \alpha \leq 48$. The matrix $B = (b_{ij})$ has the following properties:

(a) *Two entries of B in the same row or column have exactly one number β in common, $\beta > 18$.*

(b) *Two entries of B which do not stand in the same row or column have no or two numbers in common.*

(c) *Each number β with $12 < \beta \leq 48$ appears exactly six times in B .*

(d) *If $\beta > 18$, then β appears in a row or column of B or in the entries of B with fixed number α , $12 < \alpha \leq 18$, never or twice.*

For the proof of (c) one needs the theorem of Venkov.

I call a matrix B with the properties (a)–(d) a combinatorial magic square. The properties (a)–(d) seem to be very restrictive. Nevertheless there exist up to equivalence two combinatorial magic squares. One of them looks as follows:

$i \backslash j$	7	8	9	10	11	12
1	13,19,20 21,22,23	14,19,24 25,26,27	15,20,24 28,29,30	16,21,25 28,31,32	17,22,26 29,31,33	18,23,27 30,32,33
2	16,22,31 34,35,36	13,20,24 34,37,38	18,20,23 35,39,40	15,24,26 31,39,41	14,26,36 37,40,42	17,22,23 38,41,42
3	17,23,25 27,31,39	16,25,34 43,44,45	13,23,30 38,43,46	18,32,39 44,46,47	15,30,31 32,34,37	14,27,37 38,45,47
4	18,19,27 36,40,47	17,27,33 38,43,48	14,24,35 36,38,41	13,19,24 28,43,44	16,32,33 35,40,44	15,28,32 41,47,48
5	15,20,34 39,45,47	18,19,20 29,33,44	17,29,39 41,46,48	14,19,21 41,42,47	13,22,34 42,44,46	16,21,22 33,45,48
6	14,21,25 35,40,45	15,26,29 37,45,48	16,28,36 40,43,48	17,25,26 42,43,46	18,29,30 35,36,46	13,21,28 30,37,42

But it is easy to see that these magic squares do not lead to extremal codes C with $g(C) = 2$. The words $x_1, x_2, 1, a_{ij}$ with $i = 1, \dots, 6$, $j = 7, \dots, 12$ generate a doubly-even self-dual code of minimal weight 8.

In the case $g(C) = 4$ let $a \in C_{24}$ be a code word with $\dim C(a) = 4$. Without loss of generality $C(a)$ is generated by $a = \{1, \dots, 24\}$, $\{1, \dots, 12\}$, $\{1, \dots, 6, 13, \dots, 18\}$, $\{1, 2, 3, 7, 8, 9, 13, 14, 15, 19, 20, 21\}$. $C(a)$ is the $(8, 4)$ -Hamming code with the places $\{3i - 2, 3i - 1, 3i\}$, $i = 1, \dots, 8$. Let

$$u_{j_1 \dots j_8} = u_{j_1 \dots j_8}(a) \\ = |\{x \in C_{12} \mid w(x \cap \{3i - 2, 3i - 1, 3i\}) = j_i \text{ for } i = 1, \dots, 8\}|.$$

By means of the basic equations for extremal codes one can compute most of the numbers $u_{j_1 \dots j_8}$, e.g., $u_{31111111} = u_{13111111} = \dots = u_{11111113} = 3$. This provides strong restrictions to the construction of ψ (see (7)), which hopefully will allow one to classify all extremal codes C of length 48 with $g(C) = 4$.

ACKNOWLEDGMENTS. I am very grateful to B. B. Venkov from whom I learned much about unimodular lattices and self-dual codes. I would also like to thank I. Schiemann, H. Seidlitz, and C. Tamme for useful conversations and computational help.

REFERENCES

1. J. H. Conway and V. Pless, *On the enumeration of self-dual codes*, J. Combin. Theory Ser. A **28** (1980), 26–53.
2. E. Hecke, *Analytische Arithmetik der positiven quadratischen Formen*, Math. Werke, Vandenhoeck & Ruprecht, Göttingen, 1959, pp. 789–918.
3. W. C. Huffman, *Automorphisms of codes with applications to extremal doubly-even codes of length 48*, IEEE Trans. Inform. Theory **28** (1982), 511–521.
4. H. Koch, *On self-dual, doubly-even codes of length 32*, Preprint des Instituts für Mathematik der AdW der DDR P-Math-32/84, Berlin, 1984.
5. C. L. Mallows, A. M. Odlyzko, and N. J. A. Sloane, *Upper bounds for modular forms, lattices, and codes*, J. Algebra **36** (1975), 68–76.
6. C. L. Mallows and N. J. A. Sloane, *An upper bound for self-dual codes*, Inform. and Control **22** (1973), 188–200.
7. L. J. Mordell, *The definite quadratic forms in eight variables with determinant unity*, J. Math. Pures Appl. **17** (1938), 41–46.
8. H.-V. Niemeier, *Definite quadratische Formen der Dimension 24 und Diskriminante 1*, J. Number Theory **5** (1973), 142–178.
9. R. Rasala, *Split codes and the Mathieu groups*, J. Algebra **42** (1976), 422–471.
10. B. Schoeneberg, *Das Verhalten von mehrfachen Thetareihen bei Modulsubstitutionen*, Math. Ann. **116** (1939), 511–523.
11. J.-P. Serre, *Cours d'arithmétique*, Presses Univ. France, Paris, 1970.
12. B. B. Venkov, *On the classification of integral even unimodular 24-dimensional quadratic forms*, Trudy Mat. Inst. Steklov. **148** (1978), 65–76. (Russian)
13. —, *On even unimodular euclidian lattices of dimension 32*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **116** (1982), 44–55. (Russian)
14. —, Doctoral dissertation, Leningrad, 1984.
15. —, *On even unimodular extremal lattices*, Trudy Mat. Inst. Steklov. **165** (1984), 43–48. (Russian)
16. E. Witt, *Eine Identität zwischen Modulformen zweiten Grades*, Abh. Math. Sem. Univ. Hamburg **14** (1941), 323–337.

New Analytic Algorithms in Number Theory

A. M. ODLYZKO

1. Introduction. Several recently invented number-theoretic algorithms are sketched below. They all have the common feature that they rely on bounded precision computations of analytic functions. The main one of these algorithms is a new method of calculating values of the Riemann zeta function at multiple points. This method enables one to verify the truth of the Riemann Hypothesis (RH) for the first n zeros much faster than was possible before. Methods for the approximate computation of zeta and L -functions have also been shown to lead to fast algorithms for the exact computation of certain integer-valued number-theoretic functions.

Numerical computations of zeros of the zeta function cannot ever prove the RH, but if there is a counterexample to the RH, such computations might find it. This was the main motivation for most of the computations that have been carried out so far. The first few nontrivial zeros were computed by hand by Riemann [5, 23], and further computations, by hand or with the help of electromechanical devices, were carried out in the early 1900s, culminating with the verification of the RH for the first 1041 zeros (i.e., the 1041 zeros in the upper half-plane that are closest to the real axis) by Titchmarsh and Comrie [25]. After World War II, these computations were extended using electronic digital computers. The latest result in this direction is the verification of the RH for the first $1.5 \cdot 10^9$ zeros [16], a computation that involved about two months of time on a modern supercomputer, and used the same basic method for computing the zeta function that was employed by Titchmarsh and Comrie.

Computations of the zeros of the Riemann zeta function and related functions (such as Dirichlet L -functions and Epstein zeta functions) are also useful in providing evidence for and against various other conjectures, such as the Montgomery pair correlation conjecture [17, 18], and some related conjectures which predict that the zeros of $\zeta(s)$ ought to behave like eigenvalues of random hermitian matrices (whose distribution is known relatively well, since they have been studied by physicists who use them to model energy levels in many-particle systems). For a discussion of these conjectures and the numerical evidence about them, see [17, 18, 19, 20]. Another motivation for computing zeros of zeta

and L -functions comes from trying to understand better the differences between those functions, like the Riemann zeta function and Dirichlet L -function, that are expected to satisfy the RH, and those, such as the Epstein zeta functions, which are in most cases expected to violate the RH. See [2, 9] for some results on these problems. Other applications for approximate values of zeros of the zeta-function arise in disproving certain conjectures; see [10, 21].

As will be explained briefly in §2, the problem of computing an approximate value of a zero of $\zeta(s)$ and proving that this zero is on the critical line can usually be reduced to the problem of evaluating $\zeta(s)$ to medium accuracy, by which we mean computation to within $\pm t^{-c}$ for some $c > 0$. In §3, the previous methods of computing $\zeta(s)$ will be presented. The fastest of them, the Riemann-Siegel formula, requires on the order of $t^{1/2}$ operations to evaluate $\zeta(1/2 + it)$ to within $\pm t^{-c}$, say, for $t > 1$, $c > 0$. In §4 new methods for evaluating $\zeta(s)$ at multiple points are presented. The fastest of them, due to A. Schönhage and the author [22], can compute any single value of $\zeta(1/2 + it)$ with

$$T \leq t \leq T + T^{1/2}$$

to within $\pm T^{-c}$ in $T^{o(1)}$ operations (on numbers of $O(\log T)$ bits), provided a precomputation involving $T^{1/2+o(1)}$ operations is carried out beforehand and $T^{1/2+o(1)}$ storage locations are available. This method leads to an algorithm for numerically verifying the RH for the first n zeros in what is expected to be $n^{1+o(1)}$ operations, as opposed to about $n^{3/2}$ operations using the older methods. These techniques extend to the computation of other zeta and L -functions.

The computation of zeros to a specified accuracy requires only the computation of approximate values of the appropriate function. Such approximate values can be used to compute exactly various number-theoretic functions, such as $\pi(x)$, the number of primes $\leq x$. The reason for this is related to the “explicit formulas” of Guinand [8] and Weil [29], which relate arithmetic functions to zeros of the appropriate zeta and related functions. An even older formula of Landau [15] (see also [7]) says that for any fixed $y > 0$, as $T \rightarrow \infty$,

$$\sum_{0 < \operatorname{Im}(\rho) < T} e^{\rho y} = \begin{cases} -(T/2\pi) \log p + O(\log T) & \text{if } y = \log p^m, \\ O(\log T) & \text{if } y \neq \log p^m, \end{cases}$$

where p denotes a prime and m a positive integer, and ρ runs over the zeros of the zeta function. The formula above can be used to test integers for primality. This idea has occurred to many people, but unfortunately it does not seem to yield an efficient algorithm. However, modifications of such ideas can be used to compute relatively efficient functions such as $\pi(x)$, as was shown by J. C. Lagarias and the author [13, 14]. Using the latest algorithm for evaluating the zeta function, it is possible to compute $\pi(x)$ in $x^{1/2+o(1)}$ steps as $x \rightarrow \infty$, which is faster than the best known combinatorial algorithm [12], which requires around $x^{2/3}$ operations. This new algorithm for $\pi(x)$ and similar functions is sketched briefly in §5.

2. Zeta function and the numerical verification of the RH. The zeta function is defined for $\operatorname{Re}(s) > 1$ by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \quad (2.1)$$

but it can be extended by analytic continuation to an analytic function throughout $\mathbf{C} \setminus \{1\}$, and it has a first-order pole at $s = 1$. If we define, as usual,

$$\xi(s) = \pi^{-s/2} \Gamma(s/2) \zeta(s), \quad (2.2)$$

then $\xi(s)$ satisfies the functional equation

$$\xi(s) = \xi(1-s), \quad s \in \mathbf{C} \setminus \{0, 1\}. \quad (2.3)$$

Since $\xi(s)$ is real for real s , we have $\xi(\bar{s}) = \overline{\xi(s)}$, and so for $t \in \mathbf{R}$,

$$\xi(1/2 + it) = \xi(1 - (1/2 + it)) = \xi(1/2 - it) = \overline{\xi(1/2 + it)}, \quad (2.4)$$

which implies that $\xi(1/2 + it) \in \mathbf{R}$. Therefore if we find two values t_1 and t_2 such that $\xi(t_1) < 0 < \xi(t_2)$ (and the errors in the computed values of $\xi(t_1)$ and $\xi(t_2)$ are sufficiently small, so we can be certain that the two inequalities are valid), then we can conclude that there is a value t' , $t_1 < t' < t_2$, such that $\xi(1/2 + it') = 0$. In this case we know that the zero is right on the critical line, not just in some neighborhood of it.

All the numerical verifications of the RH up till now have relied on the use of the intermediate value theorem as sketched above. This theorem gives a lower bound for the number of zeros of the zeta function that are on the critical line up to some height. To assert that all of the zeros up to some height are on the critical line, one can use the principle of the argument to determine the total number of zeros up to that height and if that agrees with the number of zeros that have been located on the critical line, the verification is completed. In practice it is not even necessary to use the principle of the argument, since there is an elegant method of Turing [5, 20, 28], based on a theorem of Littlewood about the zeta function, which usually enables one to conclude that all the zeros up to some height are on the critical line based only on computations of $\xi(s)$ on that line.

It should be mentioned that the above strategy for numerically verifying the RH is not guaranteed to work, even if the RH is true, since there might be multiple zeros on the critical line. However, such zeros have not been encountered yet and are not expected to exist. In practice, using the sophisticated strategies for locating points t at which to evaluate $\xi(1/2 + it)$ that have been developed [16], it takes on average about 1.2 evaluations of $\xi(s)$ to establish that a zero is on the critical line.

The above discussion serves to reduce the problem of numerically verifying the RH to that of computing $\xi(s)$. The next section discusses that problem.

3. Methods for calculating the zeta function. For purposes of numerically verifying the RH or for computing $\pi(x)$ by the algorithm presented in §5, it is sufficient to be able to compute $\zeta(s)$ to within $\pm|s|^{-c}$ for various constants $c > 0$. For some purposes (such as those of §5) it is necessary to calculate $\zeta(s)$ for $\operatorname{Re}(s) > 1$, but to keep the presentation simple we will deal only with $\operatorname{Re}(s) = 1/2$ here. (See [22] for more details.)

The simplest method for calculating $\zeta(s)$ is to apply the Euler-Maclaurin summation formula to (2.1). This gives an analytic continuation of $\zeta(s)$ to all of $\mathbf{C} \setminus \{1\}$, and enables one to compute $\zeta(s)$ any place to arbitrary accuracy by adjusting the parameters in the formula appropriately. However, this method is not very efficient, since to compute $\zeta(1/2 + it)$ to $\pm t^{-1}$, say, already takes about t steps. Still, this was the method that was used by the first few people to publish calculations of the zeros of the zeta function.

A much more efficient method for computing $\zeta(1/2 + it)$ was known to Riemann, and was used by him to compute the first few zeros. It is one of what Hardy and Littlewood called approximate functional equations, but it has the nice feature that the error term is given by an asymptotic expansion. Discovered in Riemann's unpublished papers by Siegel, it became known as the Riemann-Siegel formula [5, 23]. It has been used for all large-scale computations of zeros of the zeta function since the 1930s. (The only other method to have been proposed so far, that of Turing [27], was meant to be used at relatively small heights, where it turned out not to be necessary.)

Let

$$\theta(t) = \arg[\pi^{-it/2}\Gamma(1/4 + it/2)], \quad (3.1)$$

$$Z(t) = \exp(i\theta(t))\zeta(1/2 + it), \quad (3.2)$$

so that

$$Z(t) = \frac{\xi(1/2 + it)}{|\pi^{-1/4 - it/2}\Gamma(1/4 + it/2)|}.$$

Also let

$$\tau = (2\pi)^{-1}t, \quad m = \lfloor \tau^{1/2} \rfloor, \quad z = 2(\tau^{1/2} - m) - 1.$$

Then the Riemann-Siegel formula [5, 6, 11, 20, 23, 25, 26] says that for any $k \geq 0$,

$$\begin{aligned} Z(t) = & 2 \sum_{n=1}^m n^{-1/2} \cos(t \log(n) - \theta(t)) \\ & + (-1)^{m+1} \tau^{-1/4} \sum_{j=0}^k \Phi_j(z) (-1)^j \tau^{-j/2} + R_k(\tau), \end{aligned} \quad (3.3)$$

where

$$R_k(\tau) = O(\tau^{-(2k+3)/4}),$$

and the $\Phi_j(z)$ are certain entire functions. Explicit bounds for the $R_k(\tau)$ are known (the best ones being due to Gabcke [6]), and bounds for the error made

in computing the $\Phi_j(x)$ are also known. Thus the entire difficulty in the computation of $Z(t)$ resides in the need to evaluate the sum of cosines in (3.3), which takes on the order of $t^{1/2}$ steps. (In the Euler-Maclaurin formula, one has to evaluate a similar sum of about t cosines.)

4. New methods for evaluating the zeta function. Evaluation of the Riemann-Siegel formula term-by-term is still the fastest known method for computing a single value of $\zeta(1/2 + it)$ to medium accuracy ($\pm t^{-c}$ for some $c > 0$) for large t . Here we will show that if values of $\zeta(1/2 + it)$ at a large set of closely spaced t 's are desired, one can obtain much faster algorithms.

Note that

$$2 \sum_{n=1}^m n^{-1/2} \cos(t \log(n) - \theta(t)) = \operatorname{Re} e^{-i\theta(t)} \sum_{n=1}^m 2n^{-1/2} e^{it \log n},$$

so that it suffices to find an efficient method to compute the complex-valued function

$$f(t) = \sum_{n=1}^m 2n^{-1/2} e^{it \log n} \quad (4.1)$$

for the values of t that are of interest. We will count the number of elementary arithmetic operations on numbers of $O(\log t)$ digits. What makes the methods below work is that there is a lot of structure in sums of the form (4.1), and if many of them are to be computed, one can exploit this structure to lower the average cost of an evaluation.

The first observation to make is that if we are interested in computing $f(t)$ (and thus $\zeta(1/2 + it)$) for $T \leq t \leq T + T^\alpha$ for some $0 < \alpha \leq 1$, say, then it suffices to compute $f(t)$ and a few similar sums at an evenly spaced grid of points, $t = t_0, t_1, \dots, t_H$, where

$$t_j = T + j\delta, \quad 0 \leq j \leq H = \lfloor T^{\alpha+\eta} \rfloor, \quad \delta = T^{-\eta},$$

for some $\eta \in (0, 1/10)$. (Eventually η will go to zero.) The reason for this is that

$$f^{(k)}(t) = \sum_{n=1}^m 2i^k n^{-1/2} (\log n)^k e^{it \log n}, \quad (4.2)$$

so that

$$|f^{(k)}(t)| \leq 5(\log m)^k m^{1/2}.$$

Therefore if we wish to compute $f(t)$ for some t , $T \leq t \leq T + T^\alpha$, to within $\pm T^{-c}$, we can use the Taylor series expansion around the nearest t_j , provided we have precomputed the values of $f^{(k)}(t_j)$ for $k \leq (c + 100)\eta^{-1}$, say. Since the derivatives $f^{(k)}(t)$ are of the same form as $f(t)$, we have reduced the problem to that of computing series of the form

$$g(t) = \sum_{n=1}^m a_n e^{it \log n} \quad (4.3)$$

at $t = t_j = T + j\delta$ for $0 \leq j \leq H$. We should note that m depends on t , $m = \lfloor (2\pi)^{-1/2} t^{1/2} \rfloor$.

The trivial way to compute $g(t)$ at each of the t_j takes about mH operations. One way to obtain a faster method is to use fast matrix multiplication methods. For simplicity, suppose that we are trying to compute $g(\delta k)$ for $0 \leq k < R^2$, where $\delta \leq 1$. Consider the matrix $C = (c_{jh})$, $0 \leq j \leq R-1$, $1 \leq h \leq R$, where

$$c_{jh} = \begin{cases} a_h \exp(ijR\delta \log h) & \text{for } h \leq \beta_j, \\ 0 & \text{for } h > \beta_j, \end{cases} \quad (4.4)$$

where the β_j will be specified later. Also, let $D = (d_{hk})$, $1 \leq h \leq R$, $0 \leq k \leq R-1$, be the matrix with

$$d_{hk} = \exp(ik\delta \log h). \quad (4.5)$$

Then the matrix $CD = (b_{jk})$, $0 \leq j, k \leq R-1$, has

$$b_{jk} = \sum_{h=1}^{\beta_j} a_h \exp(i[jR+k]\delta \log h). \quad (4.6)$$

If we select $\beta_j = \lfloor (2\pi)^{-1/2} (jR\delta)^{1/2} \rfloor$, then b_{jk} will differ from $g(\delta(jR+k))$ by only a few terms, which can be easily computed. The trivial method for multiplying C and D takes about R^3 operations, which is not much better than the ordinary $O(R^3\delta^{1/2})$ method of evaluating each sum of the form (4.6). However, if we use fast matrix multiplication methods, of which the latest allows one to multiply two $n \times n$ matrices in time $O(n^{2.479})$ [24], we obtain a method with running time about $R^{2.479}$. Since in practice we would have $\delta = R^{-o(1)}$, $R = T^{1/2+o(1)}$, this would let us compute any value of $\zeta(1/2+it)$ for $0 \leq t \leq T$ in time $T^{o(1)}$ to within $\pm T^{-c}$ after an initial precomputation that requires $\leq T^{1.24}$ operations, and would thus be more efficient than using the Riemann-Siegel formula. (This method could also be modified to work on shorter intervals, but not as efficiently.) Unfortunately fast matrix multiplication methods are completely impractical (except possibly for Strassen's original $O(n^{2.807\dots})$ method for multiplying two $n \times n$ matrices [1]), so that the above technique is of purely theoretical significance.

An algorithm that is both theoretically efficient and appears to be practical can be obtained by a different approach. Suppose that we wish to compute $g(t_j)$, $0 \leq j \leq H$, where $t_j = T + j\delta$, $0 < \delta < 1/10$, and $H \leq T^{1/2}$. Let $M = \lfloor (2\pi)^{-1/2} T^{1/2} \rfloor$, and

$$g^*(t) = \sum_{n=2}^M a_n e^{it \log n}.$$

Then $g(t_j)$ and $g^*(t_j)$ differ at most by two terms (for T large), and so it suffices to compute the $g^*(t_j)$ efficiently. Choose $N = 2^r$ so that $H < N$, $M < N$, and let

$$\omega = \exp(2\pi i/N).$$

Let

$$h(k) = \sum_{j=0}^{N-1} g^*(t_j) \omega^{jk}, \quad 0 \leq k < N.$$

The inverse Fourier transform gives

$$g^*(t_j) = \frac{1}{N} \sum_{k=0}^{N-1} h(k) \omega^{-jk}, \quad 0 \leq j \leq N-1,$$

so if we can compute all the $h(k)$ accurately, we will be able to recover the $g^*(t_j)$ in $O(N \log N)$ operations using the Fast Fourier Transform [1]. (All the numbers that come up in this procedure are of moderate size, so the required precision does not cause too much of a problem.)

The $h(k)$ can be rewritten as

$$\begin{aligned} h(k) &= \sum_{n=2}^M a_n e^{iT \log n} \sum_{j=0}^{N-1} e^{ij\delta \log n} \omega^{jk} \\ &= \sum_{n=2}^M \frac{a_n e^{iT \log n}}{1 - \omega^k e^{i\delta \log n}} = \sum_{n=2}^M \frac{-a_n e^{-i(\delta-T) \log n}}{\omega^k - e^{-i\delta \log n}}. \end{aligned}$$

(We are assuming here for simplicity that $\omega^k - \exp(-i\delta \log n)$ is not too small for any k and any n . Those n for which the above difference is small or even zero for some k have to be treated separately, cf. [22].) Thus evaluating $h(k)$ for $0 \leq k \leq N-1$ is no harder than evaluating a rational function of the type

$$h^*(z) = \sum_{n=1}^N \frac{\alpha_n}{z - \beta_n} \quad (4.7)$$

at the points $z = \omega^k$, $0 \leq k \leq N-1$, where the α_n and β_n are complex numbers with $|\beta_n| = 1$, the α_n are not too large, and $|\omega^k - \beta_n|$ is not too small for any k and any n . It is easy to show that functions of the form (4.7) can be evaluated in $O(N(\log N)^2)$ operations over some finite fields, by employing Fast Fourier Transforms methods. What is more surprising is that evaluations of such functions can also be performed in $O(N(\log N)^2)$ operations on numbers of $O(\log N)$ bits. This was shown by A. Schönhage, and is presented in [22]. The basic idea is to use Taylor series expansions of $h^*(z)$ around a small set of points. A set of functions $h_{p,q}(z)$ is defined, each of which consists of some of the terms in (4.7), such that the Taylor series of each $h_{p,q}(z)$ around a certain point converges in a relatively large region, and such that for each k , $h^*(\omega^k)$ can be written as a sum of a subset of the $h_{p,q}(\omega^k)$. Full details can be found in [22].

5. Analytic algorithms for arithmetic functions. In this section we will sketch the analytic algorithms for certain arithmetic functions that have been developed recently. For simplicity of presentation, we will restrict ourselves to the problem of computing $\pi(x)$. Full details and extensions to other functions can be found in [14].

In spite of extensive interest in computing $\pi(x)$, until recently no algorithm with proved running time of $O(x^{1-\varepsilon})$ for some $\varepsilon > 0$ was known. (See [12, 13] for historical references.) In [12] a modification of the Meissel-Lehmer algorithm, which uses combinatorial arguments and sieving, was developed. It runs in time

$x^{2/3+o(1)}$ and space $x^{1/3+o(1)}$. This algorithm has been implemented and has been used to compute values of $\pi(x)$ for various x up to $x = 4 \cdot 10^{16}$. At about the same time, a new analytic algorithm for computing $\pi(x)$ was developed [13, 14]. It relies on numerical integration of analytic transforms involving the zeta function. When one uses the Euler-Maclaurin formula to compute the zeta function, the resulting algorithm requires time $x^{2/3+o(1)}$ and space $x^{o(1)}$. With the Riemann-Siegel formula, the running time decreases to $x^{3/5+o(1)}$ and the space requirement stays at $x^{o(1)}$. With the use of the new algorithm described in §4, time drops to $x^{1/2+o(1)}$, but space required becomes $x^{1/4+o(1)}$.

The analytic algorithm for $\pi(x)$ is based on an old formula of Riemann, namely that

$$\pi(x) + \sum_{j=2}^{\infty} \frac{1}{j} \pi(x^{1/j}) = \frac{1}{2\pi i} \int_{2-i\infty}^{2+i\infty} \frac{x^s}{s} \log \zeta(s) ds \quad (5.1)$$

if $x \neq p^m$ for any prime p and $m \in \mathbf{Z}^+$. (If $x = p^m$, $(2m)^{-1}$ must be subtracted from the left side of (5.1).) One of the things which make the analytic algorithm work is that $\pi(x)$ is always an integer, and so it suffices to compute it to within $\pm 1/3$, say. The sum of the terms on the left side of (5.1) with $j \geq 2$ can easily be computed to within $\pm 1/10$ in about $x^{1/2}$ steps. (Even fewer are required if one uses the algorithm recursively.) Thus if we could compute the integral on the right side of (5.1) efficiently to within $\pm 1/10$, say, we would compute $\pi(x)$ rapidly. Unfortunately the integral in (5.1) is not even absolutely convergent, and no way has been found to compute it fast.

To overcome the problem of slow convergence of the integral in (5.1), we use a different formula, namely,

$$\sum_{p \leq x} c(p) + \sum_{\substack{p, m \\ p^m \leq x \\ m \geq 2}} \frac{1}{m} c(p^m) = \frac{1}{2\pi i} \int_{2-i\infty}^{2+i\infty} F(s) \log \zeta(s) ds, \quad (5.2)$$

where

$$c(u) = \frac{1}{2\pi i} \int_{2-i\infty}^{2+i\infty} F(s) u^{-s} ds, \quad (5.3)$$

and where this formula holds whenever $F(s)$ is a sufficiently nice function. ($F(s)$ analytic in $\operatorname{Re}(s) > 1$, $|F(s)| = |s|^{-2}$ as $|s| \rightarrow \infty$, is sufficient but not necessary, for example.) We would like to choose $F(s)$ so that (5.2) behaves like (5.1); i.e., $c(p) = 1$ for $p \leq x$ and $c(p) = 0$ for $p > x$. Unfortunately, the uniqueness theorem for Mellin transforms says that is impossible unless $F(s) = s^{-1}x^s$. We therefore select a somewhat different $F(s)$, in which the sum on the left side of (5.2) is relatively close to that on the left side of (5.1). To be precise, we select $F(s)$ so that

(i) $F(2+it) \rightarrow 0$ rapidly as $|t| \rightarrow \infty$, so that the integral in (5.2) is approximated well by a finite integral

$$\frac{1}{2\pi i} \int_{2-iT}^{2+iT} F(s) \log \zeta(s) ds; \quad (5.4)$$

(ii) Individual values of $F(s)$ can be computed rapidly, so that the integral above can be evaluated by numerical integration;

(iii) $c(p) = 1$ for primes $p \leq x - y$, $c(p) = 0$ for primes $p \geq x$, and $0 \leq c(p) \leq 1$ for primes $p \in (x - y, x)$;

(iv) Individual values of $c(k)$ can be computed efficiently.

Such $F(s)$ can be found for any y and T with the property that $yT \geq x^{1+o(1)}$; see [13, 14]. (The requirement that $yT \geq x^{1+o(1)}$ is caused by the uncertainty principle of Fourier analysis, which says that a function and its transform cannot both decrease too rapidly.) The contribution of the sum

$$\sum_{\substack{p, m \\ p^m \leq x \\ m \geq 2}} \frac{1}{m} c(p^m)$$

can be evaluated to within $\pm 1/100$ in $x^{1/2+o(1)}$ operations. The difference

$$\pi(x) - \sum_{p \leq x} c(p) = \sum_{x-y < p \leq x} (1 - c(p))$$

can be evaluated to within $\pm 1/100$ by finding all the primes in $[x - y, x]$ and computing their contribution, and this can be done in $yx^{o(1)}$ operations. Finally, the evaluation of the integral (5.4) can be reduced to evaluating $\zeta(2 + it)$ at a grid of points between $-T$ and T , spaced $x^{-o(1)}$ apart, and by the algorithm of §4 this can be done in $Tx^{o(1)}$ operations. Thus the running time of the algorithm is

$$x^{1/2+o(1)} + yx^{o(1)} + Tx^{o(1)},$$

and selecting $y = T = x^{1/2+o(1)}$, we obtain the desired running time bound.

REFERENCES

1. A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The design and analysis of computer algorithms*, Addison-Wesley, Reading, Mass., 1974.
2. E. Bombieri and D. A. Hejhal (in preparation).
3. D. Davies, *An approximate functional equation for Dirichlet L-functions*, Proc. Roy. Soc. London Ser. A **284** (1965), 224–236.
4. M. Deuring, *Asymptotische Entwicklungen der Dirichletschen L-Reihen*, Math. Ann. **168** (1967), 1–30.
5. H. M. Edwards, *Riemann's zeta function*, Academic Press, 1974.
6. W. Gabcke, *Neue Herleitung und explizite Restabschätzung der Riemann-Siegel-Formel*, Ph.D. Dissertation, Göttingen, 1979.
7. S. M. Gonek, *A formula of Landau and mean values of $\zeta(s)$* , Topics in Analytic Number Theory, S. W. Graham and J. D. Vaaler, editors, Univ. Texas Press, Austin, Texas, 1985, pp. 92–97.
8. A. P. Guinand, *A summation formula in the theory of prime numbers*, Proc. London Math. Soc. (2) **50** (1948), 107–119.
9. D. A. Hejhal, *Zeros of Epstein zeta functions and supercomputers*, Proc. Internat. Congr. Math. (Berkeley, Calif., 1986), Amer. Math. Soc., Providence, R. I., 1987.
10. A. E. Ingham, *On two conjectures in the theory of numbers*, Amer. J. Math. **64** (1942), 313–319.
11. A. Ivic, *The Riemann zeta-function*, Wiley, 1985.

12. J. C. Lagarias, V. S. Miller, and A. M. Odlyzko, *Computing $\pi(x)$: the Meissel-Lehmer method*, Math. Comp. **44** (1985), 537–560.
13. J. C. Lagarias and A. M. Odlyzko, *New algorithms for computing $\pi(x)$* , Number Theory: New York 1982, D. V. Chudnovsky, G. V. Chudnovsky, H. Cohn, and M. B. Nathanson, editors, Lecture Notes in Math., vol. 1052, Springer-Verlag, 1984, pp. 176–193.
14. —, *Computing $\pi(x)$: An analytic method*, J. Algorithms (to appear).
15. E. Landau, *Über die Nullstellen der Zetafunktion*, Math. Ann. **71** (1911), 548–564.
16. J. van de Lune, H. J. J. te Riele, and D. T. Winter, *On the zeros of the Riemann zeta function in the critical strip. IV*, Math. Comp. **46** (1986), 667–681.
17. H. L. Montgomery, *The pair correlation of zeros of the zeta function*, Proc. Sympos. Pure Math., vol. 24, Amer. Math. Soc., Providence, R. I., 1973, pp. 181–193.
18. —, *Distribution of zeros of the Riemann zeta function*, Proc. Internat. Congr. Math. (Vancouver, 1974), Vol. 1, Canad. Math. Congress, Montreal, 1975, pp. 379–381.
19. A. M. Odlyzko, *On the distribution of spacings between zeros of the zeta function*, Math. Comp. **48** (1987), 273–308.
20. —, *On the distribution of zeros of the Riemann zeta function: Conjectures and computations* (in preparation).
21. A. M. Odlyzko and H. J. J. te Riele, *Disproof of the Mertens conjecture*, J. Reine Angew. Math. **357** (1985), 138–160.
22. A. M. Odlyzko and A. Schönhage, *Fast algorithms for multiple evaluations of the Riemann zeta function* (to appear).
23. C. L. Siegel, *Über Riemanns Nachlass zur analytischen Zahlentheorie*, Quellen und Studien zur Geschichte der Math. Astr. Phys. **2** (1932), 45–80; reprinted in C. L. Siegel, *Gesammelte abhandlungen*. Vol. 1, Springer-Verlag, 1966, pp. 275–310.
24. V. Strassen, *Relative bilinear complexity and matrix multiplication* (to appear).
25. E. C. Titchmarsh, *The zeros of the Riemann zeta-function*, Proc. Roy. Soc. London **151** (1935), 234–255 and **157** (1936), 261–263.
26. —, *The theory of the Riemann zeta-function*, Oxford Univ. Press, 1951.
27. A. M. Turing, *A method for the calculation of the zeta-function*, Proc. London Math. Soc. (2) **48** (1943), 180–197.
28. —, *Some calculations of the Riemann zeta-function*, Proc. London Math. Soc. (3) **3** (1953), 99–117.
29. A. Weil, *Sur les “formules explicites” de la theorie des nombres premiers*, Comm. Sem. Math. Univ. Lund, tome supplementaire (1952), 252–265.

AT&T BELL LABORATORIES, MURRAY HILL, NEW JERSEY 07974, USA

Algebraic Groups, Hodge Theory, and Transcendence

G. WÜSTHOLZ

1. Algebraic groups and transcendence. Let G be a connected algebraic group of dimension n defined over a number field K and $\mathfrak{g} = \text{Lie } G$ its Lie algebra of invariant derivations, which has a natural structure as a K -algebra. Then we denote by $G_{\mathbb{C}}$ the Lie group obtained from G by base extension from K to \mathbb{C} . This is a complex Lie group and we have an analytic homomorphism $\exp : \mathfrak{g}_{\mathbb{C}} \rightarrow G_{\mathbb{C}}$ from $\mathfrak{g}_{\mathbb{C}} = \mathfrak{g} \otimes_K \mathbb{C}$ into $G_{\mathbb{C}}$.

Given a Lie subalgebra \mathfrak{b} of $\mathfrak{g}_{\mathbb{C}}$ we obtain by integration a Lie subgroup B of $G_{\mathbb{C}}$ which need not be closed. We say that B is defined over $\overline{\mathbb{Q}}$ if \mathfrak{b} is.

It is well known that the algebraic group G is an extension of an abelian variety A by a linear algebraic group L :

$$0 \rightarrow L \rightarrow G \rightarrow A \rightarrow 0.$$

If G is commutative then we can write L , after a base change, as a product of a power of \mathbf{G}_a , the additive group, and of \mathbf{G}_m , the multiplicative group, $L = \mathbf{G}_a^k \times \mathbf{G}_m^l$. Let us now consider the set $B(\overline{\mathbb{Q}})$ of $\overline{\mathbb{Q}}$ -valued points on B . Suppose that there exists an algebraic subgroup $H \subseteq G$ with

- (i) H defined over $\overline{\mathbb{Q}}$,
- (ii) $\dim H > 0$,
- (iii) $H_{\mathbb{C}} \subseteq B$,

where $H_{\mathbb{C}}$ is the subgroup of $G_{\mathbb{C}}$ obtained by base change. Then

$$0 \neq H(\overline{\mathbb{Q}}) = H_{\mathbb{C}}(\overline{\mathbb{Q}}) \subseteq B_{\mathbb{C}}(\overline{\mathbb{Q}}),$$

so $B_{\mathbb{C}}(\overline{\mathbb{Q}})$ is different from 0. The following theorem shows that the converse is also true.

THEOREM 1 [Wü1]. $B_{\mathbb{C}}(\overline{\mathbb{Q}}) \neq 0$ if and only if there exists an algebraic subgroup H of G such that (i), (ii), and (iii) hold.

REMARK. Actually in [Wü1] the theorem is only proved for G commutative but it is not difficult to extend the proof also to the noncommutative case.

Theorem 1 has a very straightforward corollary which proves an old conjecture of Waldschmidt and is also of some independent interest.

COROLLARY [Wü2]. *Let $\mathfrak{b} \subseteq \mathfrak{g}_{\mathbb{C}}$ be a Lie subalgebra of $\mathfrak{g}_{\mathbb{C}}$ generated over $\overline{\mathbb{Q}}$ by algebraic tangent vectors b , i.e., $b \in \exp^{-1}(G(\overline{\mathbb{Q}}))$, and $\overline{\mathfrak{b}}$ the smallest Lie subalgebra defined over $\overline{\mathbb{Q}}$ containing \mathfrak{b} . Then $\overline{\mathfrak{b}}$ is the Lie algebra of an algebraic subgroup defined over $\overline{\mathbb{Q}}$.*

Theorem 1 is actually a transcendence statement as one can see very easily. A striking example is Baker's theorem in its noneffective version. For this let $\alpha_1, \dots, \alpha_n$ be nonzero algebraic numbers.

COROLLARY 2 [Ba]. *If $\beta_1 \log \alpha_1 + \dots + \beta_n \log \alpha_n = 0$ for algebraic numbers β_1, \dots, β_n not all zero then there exist rational integers b_1, \dots, b_n not all zero such that*

$$b_1 \log \alpha_1 + \dots + b_n \log \alpha_n = 0.$$

PROOF. Let $G = G_m^n$. Then $\mathfrak{g}_{\mathbb{C}} = \mathbb{C}^n$ and

$$\exp : z = (z_1, \dots, z_n) \in \mathfrak{g}_{\mathbb{C}} \mapsto (e^{z_1}, \dots, e^{z_n}) \in G_{\mathbb{C}}.$$

Now let \mathfrak{b} be the Lie subalgebra generated by the tangent vector

$$a = (\log \alpha_1, \dots, \log \alpha_n).$$

Then $\exp(a) = (\alpha_1, \dots, \alpha_n)$ is algebraic. The dimension of $\overline{\mathfrak{b}}$ is strictly less than n since $\beta_1 z_1 + \dots + \beta_n z_n = 0$ defines a proper Lie subalgebra containing $\overline{\mathfrak{b}}$.

Corollary 1 implies now that $\overline{\mathfrak{b}}$ is algebraic and therefore the Lie algebra of an algebraic subgroup H of G of codimension at least one. But it is well known that algebraic subgroups H of G_m^n are defined in $\text{Lie } G_m^n$ by linear forms with integral coefficients. Hence there exists such a linear form $b_1 z_1 + \dots + b_n z_n$, not identically zero, which vanishes on $\overline{\mathfrak{b}}$ hence on a . This proves the corollary.

2. Sketch of the proof of Theorem 1. The proof of Theorem 1 is done by transcendence arguments and is an interplay between analytic, arithmetic, and algebraic arguments. We divide it in a natural way into a number of steps. In order to be able to give references and to avoid technical complications we restrict ourselves to the case that G is commutative.

Step 1. We first embed G into some projective space \mathbb{P}^N . For this we write G as an extension of an abelian variety A by a product of a torus G_m^k and a unipotent group which we assume to be of the form G_a^l :

$$0 \rightarrow G_m^k \times G_a^l \rightarrow G \rightarrow A \rightarrow 0.$$

Then we compactify G and obtain a homogeneous space \overline{G} on which G acts. This is described in all details in [F-W], which is based on ideas of Serre. \overline{G} now is embedded into a projective space \mathbb{P}^N for some N :

$$G \xrightarrow{i} \overline{G} \xrightarrow{\varphi} \mathbb{P}^N.$$

This gives us then the analytic functions we need for the analytic part of the proof. They are given by

$$f = \varphi \circ i \circ \exp : \mathbb{C}^n \xrightarrow{\sim} \text{Lie } G_{\mathbb{C}} \rightarrow \mathbb{P}^N.$$

These functions have order of growth at most 2. Furthermore if $f = (f_0, \dots, f_N)$ then the functions $g_i = f_i/f_0$ ($i = 1, \dots, N$) satisfy a system of differential equations, $\partial g_j/\partial z_i = h_{ij}(g_1, \dots, g_N)$ with polynomials $h_{ij}(T_1, \dots, T_N)$ for $1 \leq i \leq n$ and $1 \leq j \leq N$. These polynomials have coefficients in $\overline{\mathbf{Q}}$.

Step 2. Let ζ_1, \dots, ζ_b be coordinates on $\mathfrak{b} \simeq \mathbf{C}^b$ and $\partial/\partial\zeta_1, \dots, \partial/\partial\zeta_b$ the corresponding derivations in $\mathfrak{b} \subseteq \mathfrak{g}_{\mathbf{C}}$. Let $g \in G(\mathbf{C})$ be a complex point with $X_0(g) \neq 0$ and $V \subseteq \overline{G}$ a complete algebraic subset (not necessarily reduced); $U \subseteq \mathbf{P}^N$ the open set defined by $X_0 \neq 0$ and $V' = V \cap U$. Suppose that $g \in V(\mathbf{C})$. Then we say that g is a point of multiplicity at least T on V along B if for all $F \in I(V')$, $F = F(X_1/X_0, \dots, X_N/X_0)$, and all $\gamma = (\gamma_1, \dots, \gamma_n) \in \exp^{-1}(g)$

$$(\partial/\partial\zeta_1)^{\tau_1} \dots (\partial/\partial\zeta_b)^{\tau_b} F \circ \varphi \circ i \circ \exp(\gamma_1, \dots, \gamma_n) = 0$$

for all $0 \leq \tau_1, \dots, \tau_b$ such that $\tau_1 + \dots + \tau_b < T$. For simplicity we shall further assume that $b = n - 1$. Assume now that $g \neq 0$ is in $B_{\mathbf{C}}(\overline{\mathbf{Q}})$ and again for simplicity that g is not a torsion point. We choose the coordinates X_0, \dots, X_N such that $X_0(sg) \neq 0$ for all $s \in \mathbf{N}$. Then one constructs by the so-called Siegel Lemma for a certain triple of positive real numbers $S, T, D \gg 1$ such that $D^n \gg ST^{n-1}$, a hypersurface X on \overline{G} of degree at most D that vanishes on the set $\{sg, 0 \leq s < S\}$ to order at least T along B .

Step 3. By the so-called Schwarz Lemma one shows next that for certain real numbers S', T' with $S'T'^{n-1} \gg D^n$ the hypersurface constructed in Step 2 vanishes on the set $\{sg, 0 \leq s < S'\}$ to order at least T' .

Step 4. The vanishing statement in Step 3 can be transformed into a statement in linear algebra. It means that the coefficients of the defining equation of X have to satisfy a system of

$$S' \binom{T' + n - 1}{n - 1} \gg S'T'^{n-1}$$

linear equations, and the condition $S'T'^{n-1} \gg D^n$ just means that the number of equations is much larger than the number of coefficients, which is about D^n . In order to get a contradiction one hopes that the system has maximal rank, which would imply that all coefficients are zero; in other words $X = G$, which contradicts the definition of a hypersurface. Hence the system cannot have maximal rank. By means of some complicated machinery, the so-called “multiplicity estimates on group varieties” developed in [Ma-Wü1], [Ma-Wü2], and [Wü3], this statement is translated into statements on the group generated by g and the analytic subgroup B . The results in [Wü3] imply that the analytic subgroup B is degenerate in a certain sense; more precisely they say that B contains an algebraic subgroup H such that (i)–(iii) hold. This completes the sketch of the proof of Theorem 1.

3. Hodge theory. Let X be a smooth proper variety of dimension n over an algebraic number field K . Then there are various ways for associating with X cohomology theories.

The first one is the singular cohomology. For this let $\sigma: K \rightarrow \mathbf{C}$ be an embedding and define $X_{\mathbf{C}} := X \otimes_K \mathbf{C}$ via this embedding. Then $X_{\mathbf{C}}$ is a complex manifold, its singular cohomology $H^i(X_{\mathbf{C}}, \mathbf{Z})$ is defined, and one puts

$$H_{\mathbf{B}}^i(X) := H^i(X_{\mathbf{C}}, \mathbf{Q}) := H^i(X_{\mathbf{C}}, \mathbf{Z}) \otimes \mathbf{Q} \quad (i = 0, 1, \dots),$$

the so-called Betti-cohomology.

The de Rham cohomology is defined as follows. Let $\Omega_{X/K}^i$ ($i = 0, 1, \dots$) be the sheaf of algebraic differential i -forms. One obtains the complex

$$\Omega_{X/K}^*: \mathcal{O}_{X/K} \xrightarrow{d} \Omega_{X/K}^1 \xrightarrow{d} \Omega_{X/K}^2 \xrightarrow{d} \dots$$

Then the algebraic de Rham cohomology is defined to be

$$H_{\text{DR}}^*(X) := \mathbf{H}^*(X, \Omega_{X/K}^*),$$

where $\mathbf{H}^*(X, \Omega_{X/K}^*)$ is the hypercohomology defined in turn as follows. Let

$$\Omega_{X/K}^* \rightarrow \mathcal{J}^{*,*}$$

be a resolution of $\Omega_{X/K}^*$ by sheaves of injective \mathcal{O}_X -modules. Take its associated simple complex \mathcal{K}^* defined as $\mathcal{K}^i = \bigoplus_{a+b=i} \mathcal{J}^{a,b}$. Then

$$\mathbf{H}^*(X, \Omega_{X/K}^*) := H^*(X, \Gamma(X, \mathcal{K}^*)).$$

Then there is the Hodge-cohomology $H_{\text{Hodge}}^*(X)$ defined as

$$H_{\text{Hodge}}^i(X) := \bigoplus_{a+b=i} H^a(X, \Omega_{X/K}^b) \quad (i = 0, 1, \dots).$$

The two cohomologies $H_{\text{DR}}^*(X)$ and $H_{\text{Hodge}}^*(X)$ are related by the usual spectral sequence

$$E_1^{a,b} = H^b(X, \Omega_{X/K}^a) \Rightarrow H_{\text{DR}}^{a+b}(X)$$

and defines a filtration on $H_{\text{DR}}^*(X)$, the Hodge filtration. Now a well-known result says that

$$H^*(X, \mathbf{C}) \xrightarrow{\sim} H_{\text{DR}}^*(X) \otimes_K \mathbf{C}.$$

Using this isomorphism one defines the periods as follows: Let $e^1, \dots, e^{N_i} \in H^i(X, \mathbf{Z})$ ($i = 0, 1, \dots$) be a basis of $H^i(X, \mathbf{C})$ and e_1, \dots, e_{N_i} the dual basis in $H_i(X, \mathbf{Z})$. Furthermore let $\omega_1, \dots, \omega_{N_i} \in H_{\text{DR}}^i(X)$ ($i = 0, 1, \dots$) be a basis for $H_{\text{DR}}^i(X)$. Then

$$\omega_k = \sum_{l=1}^{N_i} \omega_{k,l} e^l \quad (k = 1, \dots, N_i).$$

The coefficients $\omega_{k,l}$ are called the periods (of weight i) and can be expressed in the usual way; namely, by integrating ω_k along $e_r \in H_i(X, \mathbf{Z})$ one finds

$$\int_{e_r} \omega_k = \sum_{l=1}^{N_i} \omega_{k,l} \cdot \int_{e_r} e^l = \sum_{l=1}^{N_i} \omega_{k,l} \delta_{r,l} = \omega_{k,r}.$$

So the complex numbers $\omega_{k,l}$ are indeed what are classically known to be periods.

The vector spaces $H^i(X, \mathbb{C})$ carry a so-called Hodge structure, i.e., they admit a canonical decomposition $H^i(X, \mathbb{C}) = \bigoplus_{a+b=i} H^{a,b}$ such that $\overline{H^{a,b}} = H^{b,a}$. They are said to be of weight i .

EXAMPLE 1. We consider an elliptic curve

$$E: y^2 = 4x^3 - g_2x - g_3, \quad g_2, g_3 \in K.$$

Then $E_{\mathbb{C}}$ is the complex torus \mathbb{C}/L for a lattice L of rank 2 that generates \mathbb{C} over the reals. Hence $H_1(X, \mathbb{Z}) = \mathbb{Z}\gamma_1 \oplus \mathbb{Z}\gamma_2$. The de Rham cohomology is obtained as follows. Let $\omega = dx/y$, $\eta = xdx/y$. Then ω is a differential of the first and η one of the second kind. They generate $H_{\text{DR}}^1(E)$ over K . The first one is everywhere holomorphic and therefore in $H^0(E, \Omega_{E/K}^1)$. The second one has two poles without residues. The periods are

$$\omega_i = \int_{\gamma_i} \omega, \quad \eta_i = \int_{\gamma_i} \eta \quad (i = 1, 2),$$

and L can be taken to be the lattice generated by ω_1 and ω_2 .

REMARK 1. It is not necessary to restrict ourselves to proper schemes. If X is arbitrary the theory of "logarithmic differentials" leads also to an analogous theory which includes differentials of the third kind.

REMARK 2. Everything extends to so-called variations of Hodge structures which arise as follows. Let S be a smooth scheme over \mathbb{C} and $f: X \rightarrow S$ a proper and smooth morphism. Then one has to replace the functors H^i by the higher direct image functors $R^i f_*$ and obtains so-called variations of Hodge structures.

REMARK 3. One often finds the situation that a Hodge structure of weight i is isomorphic to a Hodge structure coming from a Hodge structure of weight j for some j . This is where one has to introduce the so-called Tate-twists $\mathbb{Q}_{\text{B}}(1)$, $\mathbb{Q}_{\text{DR}}(1)$ defined as

$$\mathbb{Q}_{\text{B}}(1) = 2\pi i \mathbb{Q}, \quad \mathbb{Q}_{\text{DR}}(1) = K$$

and for integers m

$$H_{\text{B}}^*(X)(m) = H_{\text{B}}^*(X) \otimes \mathbb{Q}_{\text{B}}(1)^{\otimes m}, \quad H_{\text{DR}}^*(X)(m) = H_{\text{DR}}^*(X) \otimes \mathbb{Q}_{\text{B}}(1)^{\otimes m}$$

and writes $\mathbb{Q}_{\text{B}}(m) = \mathbb{Q}_{\text{B}}(1)^{\otimes m}$; and the same for $\mathbb{Q}_{\text{DR}}(m)$. $\mathbb{Q}_{\text{B}}(1)$ defines a Hodge structure of weight -2 , and its Hodge decomposition is

$$\mathbb{Q}_{\text{B}}(1) \otimes \mathbb{C} = H^{-2,0} \oplus H^{-1,-1} \oplus H^{0,-2}$$

with $H^{-2,0} = H^{0,-2} = 0$; hence $\mathbb{Q}_{\text{B}}(1) \otimes \mathbb{C} = H^{-1,-1}$. The comparison isomorphism relates the algebraic de Rham cohomology with the analytic Betti cohomology. This interplay between analytic and algebraic leads to a transcendence problem.

CONJECTURE. Let X be a smooth proper scheme of dimension n over K . Then the periods of X of weight i are either zero or transcendental.

For $i = 1$ this conjecture is due to Grothendieck [Gr]. The following theorem solves this conjecture of Grothendieck.

THEOREM 2. *Let X be a smooth proper scheme over K of dimension n . Then the periods of weight 1 are either zero or transcendental.*

The proof of this theorem uses Theorem 1 and will be published in the *Mathematische Annalen* [Wü4].

In many situations one can reduce the study of the periods of higher weight to those of weight one.

EXAMPLE 2. Let $n = 2m + 1 \geq 3$, let n, m be integral and for $\mathbf{a} = (a_1, \dots, a_r)$ with integers $a_i > 1$ let $V_n(\mathbf{a}) \subseteq \mathbf{P}^{n+r}$ be a complete intersection of dimension n of hypersurfaces of degree a_1, \dots, a_r of “niveau de Hodge 1.” This means that

- (i) $H^q(X, \Omega_X^p) = 0$ for $p + q = n$, $|q - p| > 1$,
- (ii) $H^m(X, \Omega_X^{m+1}) \neq 0$.

The list of possible $V_n(\mathbf{a})$ can be found in [Ra]. Then there exists an isomorphism of polarized Hodge structures

$$\alpha: H_{\mathbb{B}}^n(X)(m) \xrightarrow{\sim} H_{\mathbb{B}}^1(J(X)),$$

where $J(X)$ is the so-called intermediate Jacobian of X and is defined as follows. It is the complex torus

$$J(X) = H^n(X, \mathbb{Z}) \setminus H^n(X, \mathbb{C}) / H^{m+1, m},$$

where $H^n(X, \mathbb{C}) = H^{m+1, n} \oplus H^{m, m+1}$ is the Hodge decomposition. Hence $J(X) = H^{m, m+1} / H^n(X, \mathbb{Z})$. One obtains then the following corollary.

COROLLARY. *The periods of the Hodge structure $H_{\mathbb{B}}^n(X)(m)$ are either zero or transcendental.*

EXAMPLE 3. Let X be an algebraic K3-surface over K . Then again one associates with $H_{\mathbb{B}}^2(X)(1)$ via the Clifford algebra constructed from the primitive cohomology and the intersection pairing on X a certain abelian variety and again one can apply Theorem 2.

4. Vector spaces associated with rational integrals. We consider a smooth quasiprojective variety X defined over a number field K and having a K -rational point, and $\omega \in H^0(X, \Omega_{X/K}^1)$ a holomorphic differential form on X . We are interested in the periods $\int_{\gamma} \omega$, $\gamma \in H_1(X, \mathbb{Z})$. The link between these periods and Theorem 1 is given by the following result which generalizes a result of Serre [Se].

THEOREM 3 [F-W]. *Let $\omega \in \Gamma(X, \Omega_X^1)$ be a holomorphic differential form with $d\omega = 0$. Then there exists a commutative algebraic group G over K , a morphism $\varphi: X \rightarrow G$, and an $\alpha \in \mathfrak{g}^* = \Gamma(G, \Omega_G^1)^{\text{const.}}$, $\mathfrak{g} := \text{Lie } G$, such that $\omega = \varphi^* \alpha$.*

The proof of this theorem uses the construction of the so-called generalized Albanese variety. If one combines this result and Theorem 1 one obtains the following theorem.

THEOREM 4 [Wü4]. *The numbers $\int_{\gamma} \omega$ ($\gamma \in H_1(X, \mathbf{Z})$) are either zero or transcendental.*

It is a natural question to look at vector spaces generated by numbers $\int_{\gamma} \omega$ ($\gamma \in H_1(X, \mathbf{Z})$, $\omega \in H^0(X, \Omega_{X/K}^1)$). Since in general $H^0(X, \Omega_{X/K}^1)$ is not finite-dimensional, one has to look at canonical finite-dimensional subspaces of this space.

Special cases of questions of this type have been treated by various authors (Siegel, Schneider, Baker, Coates, Masser). But they all dealt with low-dimensional algebraic groups. Till now we have not been able to work out a general result on this question depending intrinsically on the cohomology of X . But in important special cases we can give explicit results. For this let Y be a smooth projective variety and $D \subseteq Y$ an ample divisor, both defined over some number field K . Then put $X = Y \setminus D$. In this case we have a surjection

$$\pi: H^0(X, \Omega_{X/K}^1) \rightarrow H_{\text{DR}}^1(Y)$$

by a result of Hodge and Atiyah [H-A]. Therefore we can find a basis for $H_{\text{DR}}^1(Y)$ (up to exact forms) from forms in $H^0(X, \Omega_{X/K}^1)$. We take therefore a section of π , i.e., a subspace V of $H^0(X, \Omega_{X/K}^1)$. Then we denote by $V(X)$ the vector space generated by $1, 2\pi i, \int_{\gamma} \omega$ ($\gamma \in H_1(X, \mathbf{Z})$, $\omega \in V$). Let $\text{Alb}(Y)$ be the Albanese variety of Y . Then up to isogeny we have

$$\text{Alb}(Y) = A_1^{n_1} \times \cdots \times A_k^{n_k}$$

for pairwise nonisogenous simple abelian varieties A_1, \dots, A_k defined over \overline{K} . Then we have the following result.

THEOREM 5 [Wü4]. *We have*

$$\dim V = 2 + 4 \sum_{i=1}^k \frac{(\dim A_i)^2}{\dim(\text{End } A_i)}.$$

One can for example apply this result in the case that Y is a Fermat curve.

EXAMPLE 4. Let X be the Fermat curve of degree N given by

$$X_0^N + X_1^N + X_2^N = 0$$

where $N \geq 3$. Then a basis for $H_{\text{DR}}^1(X)$ is given by the differentials $(x, y$ affine coordinates)

$$\omega_{a,b} = x^{Na-1} y^{Nb-N} dx \quad (\text{first kind})$$

and

$$\eta_{a,b} = x^{N-Na-1} y^{-Nb} dy \quad (\text{second kind})$$

where a and b are rational numbers satisfying $0 < a, b < 1$, $a+b < 1$, $Na, Nb \in \mathbf{Z}$. The periods are given by numbers of the form

$$\int_{\gamma} \omega_{a,b} = c(a, b)B(a, b), \quad \int_{\gamma} \eta_{a,b} = c'(a, b)B(1-a, 1-b)$$

where $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ when defined and $c(a, b), c'(a, b) \in \overline{\mathbb{Q}}$. It is possible to decompose explicitly the Jacobian $J(N)$ of this Fermat curve and to obtain the dimension of the vector space generated by the numbers $1, 2\pi i, B(a, b), B(1-a, 1-b)$ for the given range of a, b . In particular one obtains the following corollary.

COROLLARY. *For rational numbers a, b with $0 < a, b < 1$, $a + b < 1$ the numbers $1, \pi, B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b), \pi/B(a, b)$ are $\overline{\mathbb{Q}}$ -linearly independent.*

As an application of this corollary to a problem of Lang let X be a smooth projective curve over $\overline{\mathbb{Q}}$ of genus $g > 1$. Then the universal covering space is the open unit disc B_1 . Let $x \in X(\overline{\mathbb{Q}})$ be a $\overline{\mathbb{Q}}$ -rational point and $\varphi_1: B_1 \rightarrow X_{\mathbb{C}}$ a covering map with $\varphi_1(0) = x$. After a homothety of B_1 into B_r for some $r \in \mathbb{C}^\times$ one can get a new covering map $\varphi_r: B_r \rightarrow X_{\mathbb{C}}$ such that $\varphi_r(0) = x$ and $\varphi'_r(0) \in \overline{\mathbb{Q}}$.

THEOREM 6 [W-W]. *Let Γ_1 be a torsion free Fuchsian group of finite index in a triangular group $\Delta_1 \subset \text{Aut } B_1$ with signature (p, q, s) and 0 an elliptic flux point of order $p > 1$ of Δ_1 , $X = \Gamma_1 \setminus B_1$. Then r is transcendental.*

REFERENCES

- [Ba] A. Baker, *Linear forms in logarithms of algebraic numbers*. I, *Mathematika* **13** (1966), 204–216.
- [F-W] G. Faltings and G. Wüstholz, *Einbettungen kommutativer algebraischer Gruppen und einige ihrer Eigenschaften*, *J. Reine Angew. Math.* **354** (1984), 175–205.
- [Gr] A. Grothendieck, *On the de Rham cohomology of algebraic varieties*, *Inst. Hautes Études Sci. Publ. Math.* **29** (1966), 95–103.
- [H-A] W. V. D. Hodge and M. F. Atiyah, *Integrals of the second kind on an algebraic variety*, *Ann. of Math.* (2) **62** (1955), 56–91.
- [Ma-Wü1] D. W. Masser and G. Wüstholz, *Zero estimates on group varieties*. I, *Invent. Math.* **64** (1981), 489–516.
- [Ma-Wü2] ———, *Zero estimates on group varieties*. II, *Invent. Math.* **80** (1985), 233–267.
- [Ra] M. Rappoport, *Complément à l'article de P. Deligne "La conjecture de Weil pour les surfaces K3"*, *Invent. Math.* **15** (1972), 227–236.
- [Se] J.-P. Serre, *Groupes algébriques et corps de classes*, Hermann, Paris, 1959.
- [W-W] J. Wolfart and G. Wüstholz, *Der Überlagerungsradius gewisser algebraischer Kurven und die Werte der Betafunktion an rationalen Stellen*, *Math. Ann.* **273** (1985), 1–15.
- [Wü1] G. Wüstholz, *Algebraische Punkte auf analytischen Untergruppen algebraischer Gruppen*, *Ann. of Math.* (2) (to appear).
- [Wü2] ———, *Some remarks on a conjecture of Waldschmidt*, *Approximations Diophantennes et Nombres Transcendants*, *Progr. Math.*, vol. 31, Birkhäuser, 1983, pp. 329–336.
- [Wü3] ———, *Multiplicity estimates on group varieties*, *Ann. of Math.* (2) (to appear).
- [Wü4] ———, *Periods of rational integrals*, *Math. Ann.* (to appear).

BERGISCHE UNIVERSITÄT GESAMTHOCHSCHULE WUPPERTAL, 5600 WUPPERTAL 1,
FEDERAL REPUBLIC OF GERMANY

Manifolds of Nonpositive Sectional Curvature and Manifolds without Conjugate Points

WERNER BALLMANN

Let M be a complete Riemannian manifold. For a tangent vector v of M , denote by γ_v the geodesic with initial velocity v . The geodesic flow g^t of M is defined by $g^t(v) = \dot{\gamma}_v(t)$. The geodesic flow acts on the unit tangent bundle SM of M , and it leaves the natural measure of SM —the Liouville measure—invariant. Denote by μ the normalized Liouville measure.

The synthetic characterization of manifolds without conjugate points says that M does not have conjugate points if and only if for each pair of points $p \neq q$ in the universal covering space \tilde{M} of M there is a unique geodesic passing through p and q . This implies that \tilde{M} is diffeomorphic to \mathbf{R}^n , $n = \dim(M)$, and hence that M is a $K(\pi, 1)$.

If the sectional curvature K of M is nonpositive, and if γ_1 and γ_2 are unit speed geodesics in \tilde{M} , then the function $d(t) = \text{dist}(\gamma_1(t), \gamma_2(t))$ is convex. In particular, M does not have conjugate points. Moreover, if $d(t)$ is bounded, then γ_1 and γ_2 bound a common flat strip, see [EO].

1. Measure entropy of g^t and structure of $\pi_1(M)$. Suppose now that M is compact and has nonpositive sectional curvature. For $v \in SM$, let $W = W(v) \subset T_v SM$ be the tangent space to the C^1 -submanifold of SM consisting of vectors asymptotic to v , with footpoint on the horosphere determined by v . Then the limit

$$\lim_{t \rightarrow \infty} \frac{1}{t} \ln |\det(dg^t|_W)| := \chi(v) \quad (1.1)$$

exists for almost every v , see [Oe], and the measure theoretic entropy of g^t (with respect to the measure μ) is given by $h_\mu = -\int_{SM} \chi(v) d\mu(v)$, see [P1, P2]. Denote by $U(v)$ the second fundamental form, in the footpoint of v , of the horosphere determined by the vector $v \in SM$. In our normalization, $U(v)$ is negative semidefinite, and it acts on the orthogonal complement E_v of v in

Research supported by National Science Foundation Grant DMS-85-03742.

$T_p M$, $p = \text{foot}(v)$. Pesin proved

$$h_\mu = - \int_{SM} \text{tr}(U(v)) d\mu(v), \quad (1.2)$$

see [P2]. Indeed, the space W in (1.1) consists of pairs of vectors $(Y, Z) \in E_v \oplus E_v \subset T_v SM$ such that the Jacobi field $J(t)$ along $\gamma_v(t)$ determined by $J(0) = Y$ and $J'(0) = Z$ is monotonically not increasing. Then $Z = U(v) \cdot Y$. If $-a^2$ is a lower bound for the sectional curvature of M , then $\|Z\| \leq a\|Y\|$, see [E1], and hence

$$\chi(v) = \lim_{t \rightarrow \infty} \frac{1}{t} \ln |J_2(t) \wedge \cdots \wedge J_n(t)|,$$

where $J_2(t), \dots, J_n(t)$ are Jacobi fields, as above, such that $J_2(0), \dots, J_n(0)$ are a basis of E_v . Differentiation yields

$$\chi(v) = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \text{tr}(U(g^t v)) dt.$$

Applying the Birkhoff ergodic theorem, cf. for example [AA], we obtain the formula (1.2).

The horosphere determined by $g^t v$ is parallel to the horosphere determined by v , and hence the family $U(v)$, $v \in SM$, satisfies the *Riccati equation*

$$U' + U^2 + S = 0, \quad (1.3)$$

where $U'(v)$ denotes the covariant derivative of U along γ_v and $S(v) \cdot X := R(X, v)v$. The Riccati equation relates U to the curvature of M .

If the curvature of M is negative, then S and U are negative definite and hence invertible, and we get $U'U^{-1} + U + SU^{-1} = 0$. Therefore

$$(\ln(\det(U)))' + \text{tr}(U) + \text{tr}(SU^{-1}) = 0.$$

Since g^t preserves μ we obtain

$$- \int_{SM} \text{tr}(U(v)) d\mu(v) = \int_{SM} \text{tr}(S(v)U^{-1}(v)) d\mu(v).$$

Now recall that $\text{tr}(AA^t) \geq 0$ for every matrix A , with equality if and only if $A = 0$. Applying this we get

$$\begin{aligned} 0 &\leq \text{tr}((\sqrt{-U} - \sqrt{-S}\sqrt{-U}^{-1}) \cdot (\sqrt{-U} - \sqrt{-S}\sqrt{-U}^{-1})^t) \\ &= \text{tr}(-U - 2\sqrt{-S} + \sqrt{-S}(-U)^{-1}\sqrt{-S}) \\ &= -\text{tr}(U) + \text{tr}(SU^{-1}) + 2\text{tr}\sqrt{(-S)} \end{aligned}$$

since $\sqrt{-S}$ and $\sqrt{-U}$ are symmetric. Hence

$$h_\mu \geq - \int_{SM} \text{tr}(\sqrt{-S(v)}) d\mu(v). \quad (1.4)$$

Equality implies $S = U^2$, and hence $U' = 0$ by the Riccati equation. So U^2 and therefore S are parallel along geodesics, and it follows easily that M is a locally symmetric space (of noncompact type and rank one). The inequality (1.4) (with

the equality discussion) is due to Osserman and Sarnak [OS], and it contains all previously known lower estimates for h_μ for negatively curved manifolds. The argument above is actually a simplification of the arguments in [OS] and is due to Wojtkowski. It can be shown that (1.4) is also true under the weaker assumption of nonpositive sectional curvature, see [BW]. The main difficulty in extending the above argument lies in the fact that U is not necessarily invertible.

For any compact Riemannian manifold M , the action of $\pi_1(M)$ on \tilde{M} admits a compact fundamental domain $\Delta \subset \tilde{M}$ with piecewise smooth boundary. Using a result of Følner, Brooks [Br] and Gromov [GLP] showed that $\pi_1(M)$ is amenable if and only if

$$I_\Delta := \inf(\text{vol}(\partial A)/\text{vol}(A)) = 0 \quad (1.5)$$

where the infimum is taken over all domains A in \tilde{M} of the form $A = \bigcup_{\varphi \in B} \varphi \cdot \Delta$, $B \subset \pi_1(M)$ finite. Recall that solvable groups are amenable and that $\pi_1(M)$ has exponential growth if it is not amenable.

Now assume again that M has nonpositive sectional curvature. In a recent paper, Anderson proves the inequality $I_\Delta \geq h_\mu$, see [An]. As a corollary of this and (1.4) we obtain that M is flat if $\pi_1(M)$ is amenable. This latter fact is already mentioned in [GLP], and it strengthens results of Avez [Av], Gromoll-Wolf [GW], and Yau [Ya]. Define

$$I_M = \inf\{\text{vol}(\partial A)/\text{vol}(A) \mid A \text{ compact domain in } \tilde{M}\}.$$

Then $I_M \leq I_\Delta$, but it is not yet known whether $I_M \geq h_\mu$.

We now relax the assumption that M has nonpositive sectional curvature and require only that M has no conjugate points. Using Jacobi fields one can still define the operators $U(v)$. Namely, if $Y \in E_v$ and J_T denotes the unique Jacobi field along γ_v such that $J_T(0) = Y$ and $J_T(T) = 0$, then $U(v) \cdot Y = \lim_{t \rightarrow \infty} J'_T(t)$. Since these limits exist for each $v \in SM$, the family of $U(v)$, $v \in SM$, is measurable in v . Moreover, we still have $h_\mu = \int_{SM} \text{tr}(U) d\mu$, see [FM], and $I_\Delta \geq h_\mu$, see [An]. In particular, $h_\mu = 0$ if $\pi_1(M)$ is amenable. The problem arises whether M is flat if $h_\mu = 0$. This is unknown. Combining results of Pesin [P3] and Knieper [Kn], we find that a compact surface without conjugate points, such that $U(v)$ is continuous in v , is either flat or $h_\mu > 0$ and the geodesic flow is ergodic. It was claimed by E. Hopf that $U(v)$ is continuous in v on such a surface [Ho]; however, a counter example to this is given in [BBB].

2. Rank of manifolds of nonpositive sectional curvature. Throughout this section we assume that M is a complete Riemannian manifold of nonpositive sectional curvature. For $v \in SM$, define $J^P(v)$ to be the space of parallel Jacobi fields along γ_v and set

$$\text{rank}(v) = \dim J^P(v) \quad \text{and} \quad \text{rank}(M) = \min\{\text{rank}(v) \mid v \in SM\}.$$

If the sectional curvature of M is negative, then the rank of M is one. A surface of nonpositive curvature has rank one if and only if it is not flat. It is also easy to see that the above notion of rank coincides with the usual one in the case that M is locally symmetric. Note that $\text{rank}(M) \geq 2$ if \tilde{M} is a Riemannian product.

A manifold of rank one resembles in many ways a manifold of negative curvature.

THEOREM 2.1. *Suppose M has rank one.*

- (a) *If $\text{vol}(M) < \infty$, then the geodesic flow of M is topologically transitive.*
- (b) *If M is compact, then the geodesic flow of M is ergodic.*

The second part of this theorem, proved in [BB] and independently in [Bu], generalizes the result of Anosov that the geodesic flow on a compact manifold of negative curvature is ergodic [Ao]. It also generalizes a result of Pesin [P3], namely that the geodesic flow on a compact surface of nonpositive curvature is ergodic if and only if the surface is not flat.

The first part of the above theorem is proved in [B1] under the at-first-sight weaker assumption that M has a geodesic which does not bound a totally geodesic immersed flat half plane (and $\text{vol}(M) < \infty$). However, it was shown in [Bu] that this implies that M has rank one. Burns's result was a first step towards

THEOREM 2.2 [BBE]. *Suppose $\text{vol}(M) < \infty$ and $\text{rank}(M) = k \geq 2$. Then every geodesic in M is contained in a k -flat, that is, a totally geodesic and isometrically immersed Euclidean space of dimension k .*

This is one of the basic results for the investigations in [BBE] and [BBS]. Another basic result is the Angle Lemma [BBE], in which the further hypothesis $K \geq -a^2$ is needed. To formulate the Angle Lemma, call a vector $v \in \tilde{SM}$ *regular* if $\text{rank}(v) = \text{rank}(M) = k$, and denote by $F(v)$ the (unique) k -flat containing v . The Angle Lemma asserts that for a dense set of regular vectors v in \tilde{SM} (namely the *uniformly recurrent* regular vectors) there is an open neighbourhood of v in $S_p F(v)$, $p = \text{foot}(v)$, such that any w' asymptotic to some $w \in U$ is regular and $F(w') = F(v')$, where v' denotes the unique vector asymptotic to v such that $\text{foot}(v') = \text{foot}(w')$. This leads to the definition of *Weyl chambers* in \tilde{SM} (and SM) [BBS]. We say that two unit vectors v and w tangent to a flat F at a point $p \in \tilde{M}$ belong to the same Weyl chamber if $F(v(q)) = F(w(q))$ for all $q \in \tilde{M}$ such that $v(q)$ and $w(q)$ are regular. Here $v(q)$ and $w(q)$ denote the vectors at q asymptotic to v and w respectively. The Weyl chamber of v is denoted by $C(v)$. In the case that M is a symmetric space, this definition of Weyl chamber corresponds to the usual definition of Weyl chamber. Weyl chambers define an equivalence relation on the set of regular vectors.

The Angle Lemma shows that there are v which are contained in the interior of $C(v)$ if $\text{rank}(M) = k \geq 2$. On an open, dense, and g^t -invariant subset O of SM , Weyl chambers depend continuously on v , and $C(w)$ is locally isometric to $C(v)$ if w is sufficiently close to v according to the *Rigidity Lemma* [BBS]. If v is regular, then the map $w \rightarrow w(q)$ gives rise to an isometry of $C(v)$ and $C(v(q))$ for all q such that $v(q)$ is regular. In particular, the function

$$\Phi: O \rightarrow \mathbf{R}; \quad \Phi(v) = \min\{\angle(v, w) \mid w \in \partial C(v)\}$$

is a continuous integral for the geodesic flow on O . Clearly Φ is not constant if $\text{rank}(M) \geq 2$ and thus we get the following result.

THEOREM 2.3 [BBS]. *Suppose $\text{vol}(M) < \infty$, $K \geq -a^2$, and $\text{rank}(M) = k \geq 2$. Then the geodesic flow of M is not topologically transitive.*

One can actually construct $k - 1$ independent differentiable first integrals for g^t [BBS]. Using a generalized version of Anosov's *Closing Lemma* [An], one can also show that a dense set of vectors $v \in SM$ is contained in an immersed totally geodesic flat k -torus [BBS]. We now come to the main result about manifolds of higher rank.

THEOREM 2.4. *Suppose $\text{vol}(M) < \infty$ and $K \geq -a^2$. Then \tilde{M} is a space of rank one or a symmetric space or a Riemannian product of such spaces.*

This theorem is proved in [B2] and [BS1]. Both proofs rely on the previous results in [BBE] and [BBS], but apart from that they are completely different.

The proof in [B2] uses Berger's theorem that \tilde{M} is a symmetric space of higher rank if it is irreducible and if the holonomy group at some point $p \in \tilde{M}$ does not act transitively on $S_p\tilde{M}$, see [Be, Si]. Consider the function Φ as above. It turns out that Φ is a Lipschitz function on O and thus has a continuous extension to a Lipschitz function $\Phi: S\tilde{M} \rightarrow \mathbf{R}$. (Recall that O is dense in $S\tilde{M}$.) Moreover, Φ is constant along the C^1 -submanifolds $A^+(v) = \{v(q) \mid q \in \tilde{M}\}$ of $S\tilde{M}$. One can also show that Φ is constant along the C^1 -submanifolds $A^-(v) = -A^+(-v)$, $v \in S\tilde{M}$. This implies that the directional derivative of Φ in the direction of any $X \in T_v A^+(v)$ or $X \in T_v A^-(v)$, $v \in S\tilde{M}$, exists and vanishes. Since the families $A^+(v)$ and $A^-(v)$, $v \in S\tilde{M}$, foliate $S\tilde{M}$ and depend continuously on v in the C^1 -topology, and since $T_v A^+(v) + T_v A^-(v)$ contains the horizontal subspace of $T_v S\tilde{M}$ for every $v \in S\tilde{M}$, it follows that the directional derivative of Φ in any horizontal direction exists and vanishes. Thus $\Phi(w(t)) = \Phi(w(0))$ if $w(t)$ is a parallel vector field along a curve in M . If $\text{rank}(M) \geq 2$, then Φ is not constant on $S_p\tilde{M}$, where $p \in M$ is arbitrary. Since the orbits of the holonomy group at p are contained in the level surfaces of Φ , they do not fill $S_p\tilde{M}$. Hence \tilde{M} is either reducible or a symmetric space of higher rank according to Berger's theorem.

The proof of (2.4) by Burns and Spatzier [BS1] relies on an extension of the theory of Tits buildings to a theory of Tits buildings with topology, cf. [BS2]. Assume for simplicity that \tilde{M} is irreducible and $\text{rank}(\tilde{M}) \geq 2$. The Tits building T associated to \tilde{M} is given by the points at infinity, $\tilde{M}(\infty)$, together with the subsets defined by the Weyl chambers. Each element of the fundamental group Γ of \tilde{M} induces a continuous automorphism of T . The group of all continuous automorphisms of T is a centerless Lie group G of noncompact type, and Γ is a discrete lattice in G . Vice versa, the Tits building is also determined by the group G ; that is, T is also the Tits building associated to the symmetric space G/K , where K is a maximal compact subgroup of G . Given a point $p \in \tilde{M}$, the geodesic flip about p defines an involutive automorphism of T (not necessarily an isometry of \tilde{M} a priori). Such involutive automorphisms correspond to geodesic

symmetries in G/K . Thus we can associate to p the fixed point $\psi(p)$ of that geodesic symmetry. This idea is due to Gromov, and he proved that ψ is an isometry up to a scaling factor, see [BGS]. Note that ψ is Γ -invariant, and hence (2.4) follows.

It is remarkable that the rank of M is determined by the algebraic structure of $\pi_1(M)$ if $\text{vol}(M) < \infty$ and $K \geq -a^2$, see [BE]. This leads to a characterization of irreducible locally symmetric spaces of noncompact type of higher rank in terms of properties of $\pi_1(M)$, cf. [BE].

From now on we assume that M is compact. By Theorem 2.2, $\text{rank}(M)$ is the maximal number k such that every geodesic of M is contained in a k -flat. Another natural notion of rank for M is $\text{Rank}(M) = \max\{k \mid M \text{ contains a } k\text{-flat}\}$. Once again, this coincides with the usual notion of rank for locally symmetric spaces. Recall that M contains a totally geodesic and isometrically immersed k -torus if $\pi_1(M)$ contains a free abelian subgroup of rank k , see [GW] and [LY], and thus

$$\text{Rank}(M) \geq \max\{k \mid \pi_1(M) \text{ contains a free abelian subgroup of rank } k\}. \quad (2.5)$$

Equality should hold, at least if M is analytic. So far the following is known.

THEOREM 2.6 [AS]. *Rank(M) is the maximal number k such that there is a quasi-isometry (with respect to word norms) of \mathbb{Z}^k into $\pi_1(M)$.*

Thus $\text{Rank}(M)$ is determined by $\pi_1(M)$.

We say that M satisfies the Visibility Axiom if any two distinct points in the ideal boundary of \tilde{M} can be joined by a geodesic [EO]. Eberlein showed that M satisfies the Visibility Axiom if and only if $\text{Rank}(M) = 1$, see [E2]. Therefore equality in (2.5) would show that M satisfies the Visibility Axiom if and only if $\pi_1(M)$ satisfies the Preismann property that every abelian subgroup $\neq \{e\}$ is infinite cyclic.

REFERENCES

- [An] M. Anderson, *Geodesic flow on manifolds without conjugate points*, Preprint, Pasadena, 1986.
- [AS] M. Anderson and V. Schroeder, *Existence of flats in manifolds of nonpositive curvature*, Preprint, Pasadena-Basel, 1985.
- [Ao] D. Anosov, *Geodesic flows on closed Riemannian manifolds with negative curvature*, Proc. Steklov Inst. Math., vol. 90, Amer. Math. Soc., Providence, R. I., 1969.
- [AA] V. Arnold and A. Avez, *Problèmes ergodiques de la mécanique classique*, Gauthier-Villars, Paris, 1967.
- [Av] A. Avez, *Variétés Riemanniennes sans points focaux*, C. R. Acad. Sci. Paris Sér A-B **270** (1970), 188–191.
- [B1] W. Ballmann, *Axial isometries of manifolds of non-positive curvature*, Math. Ann. **259** (1982), 131–144.
- [B2] ———, *Nonpositively curved manifolds of higher rank*, Ann. of Math. **122** (1985), 597–609.
- [BB] W. Ballmann and M. Brin, *On the ergodicity of geodesic flows*, Ergodic Theory Dynamical Systems **2** (1982), 311–315.
- [BBB] W. Ballmann, M. Brin, and K. Burns, *On surfaces with no conjugate points*, Preprint, College Park, 1986.

- [BBE] W. Ballmann, M. Brin, and P. Eberlein, *Structure of manifolds of nonpositive curvature. I*, Ann. of Math. **122** (1985), 171–203.
- [BBS] W. Ballmann, M. Brin, and R. Spatzler, *Structure of manifolds of nonpositive curvature. II*, Ann. of Math. **122** (1985), 205–235.
- [BE] W. Ballmann and P. Eberlein, *Fundamental group of manifolds of nonpositive curvature*, J. Differential Geom. (to appear).
- [BGS] W. Ballmann, M. Gromov, and V. Schroeder, *Manifolds of nonpositive curvature*, Birkhäuser, Boston-Basel-Stuttgart, 1985.
- [BW] W. Ballmann and M. Wojtkowski, *Estimates for the measure entropy of geodesic flows* (in preparation).
- [Be] M. Berger, *Sur les groupes d'holonomie homogène des variétés à connexion affine et des variétés Riemanniennes*, Bull. Soc. Math. France **83** (1955), 279–330.
- [Br] R. Brooks, *The fundamental group and the spectrum of the Laplacian*, Comment. Math. Helv. **56** (1981), 581–598.
- [Bu] K. Burns, *Hyperbolic behaviour of geodesic flows on manifolds with no focal points*, Ergodic Theory Dynamical Systems **3** (1983), 1–12.
- [BS1] K. Burns and R. Spatzier, *Manifolds of nonpositive curvature and their buildings*, Preprint, College Park-Stony Brook, 1985.
- [BS2] —, *Classification of a class of topological Tits buildings*, Preprint, College Park-Stony Brook, 1985.
- [E1] P. Eberlein, *When is a geodesic flow of Anosov type? I*, J. Differential Geom. **8** (1973), 437–463.
- [E2] —, *Geodesic flow in certain manifolds without conjugate points*, Trans. Amer. Math. Soc. **167** (1972), 151–170.
- [EO] P. Eberlein and B. O'Neill, *Visibility manifolds*, Pacific J. Math. **46** (1973), 45–109.
- [FM] A. Freire and R. Mañé, *On the entropy of the geodesic flow in manifolds without conjugate points*, Invent. Math. **69** (1982), 375–392.
- [GW] D. Gromoll and J. Wolf, *Some relations between the metric structure and the algebraic structure of the fundamental groups in manifolds of nonpositive curvature*, Bull. Amer. Math. Soc. **77** (1971), 545–552.
- [GLP] M. Gromov, J. Lafontaine, and P. Pansu, *Structure métriques pour les variétés Riemanniennes*, Cedric: Fernand Nathan, Paris, 1981.
- [Ho] E. Hopf, *Closed surfaces without conjugate points*, Proc. Nat. Acad. Sci. U.S.A. **34** (1948), 47–51.
- [Kn] G. Knieper, *Mannigfaltigkeiten ohne konjugierte Punkte*, Inaugural Dissertation, Bonn, 1985.
- [LY] H. B. Lawson and S.-T. Yau, *Compact manifolds of nonpositive curvature*, J. Differential Geom. **7** (1972), 211–228.
- [Oe] V. Oseledec, *A multiplicative ergodic theorem*, Trans. Moscow Math. Soc. **19** (1968), 197–231.
- [OS] R. Osserman and P. Sarnak, *A new curvature invariant and entropy of geodesic flows*, Invent. Math. **77** (1984), 455–462.
- [P1] Ja. Pesin, *Characteristic Lyapunov exponents and smooth ergodic theory*, Russian Math. Surveys **32** (1977), no. 4, 55–114.
- [P2] —, *Equations for the entropy of a geodesic flow on a compact Riemannian manifold without conjugate points*, Math. Notes **24** (1978), 796–805.
- [P3] —, *Geodesic flows on closed Riemannian manifolds without focal points*, Math. USSR-Izv. **11** (1977), 1195–1228.
- [Si] J. Simons, *On the transitivity of holonomy systems*, Ann. of Math. **76** (1962), 213–234.
- [Ya] S.-T. Yau, *On the fundamental group of compact manifolds of non-positive curvature*, Ann. of Math. **93** (1971), 579–585.

UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND 20742, USA

MATHEMATISCHES INSTITUT, UNIVERSITÄT BONN, D 5300 BONN 1, FEDERAL REPUBLIC OF GERMANY

Index Theorem and the Heat Equation

JEAN-MICHEL BISMUT

The purpose of this paper is to review the recent developments in the heat equation proofs of the Atiyah-Singer Index Theorem for Dirac operators.

Let us briefly recall that if D_+ is half a Dirac operator and if D_- is its adjoint, the starting point of the method is the McKean-Singer formula for the index $\text{Ind } D_+$ [MS]:

$$[\text{Ind } D_+ = \text{Tr}[\exp(-tD_-D_+/2)] - \text{Tr}[\exp(-tD_+D_-/2)], \quad t > 0. \quad (0.1)$$

As $t \downarrow 0$, the right-hand side has an expansion starting with negative powers of t , and the problem is to show that in certain situations, when expressing the traces using kernels, even locally

- no negative powers of t arise;
- the constant term coincides with the local Atiyah-Singer polynomial.

After the pioneering papers of Patodi [P1, P2], Gilkey [Gi1] and Atiyah-Bott-Patodi [ABP] established that this is indeed the case for algebraic reasons. The method is indirect:

- an algebraic argument gives the general form of the local terms arising in the expansion, and then excludes negative powers of t ;
- the zero order term is calculated using a similar classification argument, and also explicit computations on examples.

This approach has been extended by Patodi [P3] and Gilkey [Gi2] to include the fixed point formulas of Atiyah-Bott [AB1] and Atiyah-Singer [AS], and it is fully described in Gilkey's recent book [Gi3].

By using arguments based on supersymmetry considerations, physicists Witten [W1], Alvarez-Gaumé [A1], and Friedan-Winney [FW] strongly suggested that a direct proof of the local Index Theorem could be given, which would altogether prove the local cancellations and identify the local integrand by brute force.

That this is indeed possible has been proved by Getzler [Ge1, Ge2] for the Index Theorem, by Bismut [B1] and Berline-Vergne [BV2] for the Index Theorem and for the Lefschetz fixed point formulas. The proofs of Getzler are based on the asymptotic representation of heat kernels on supermanifolds [Ge1] and

also on adequate rescaling in time, space, and Clifford variables [Ge1, Ge2]. Our proofs [B1] use a probabilistic asymptotic representation of the heat kernel [B5], together with certain stochastic area formulas of P. Lévy [Le]. The proofs of Berline-Vergne [BV2] are of group-theoretic nature.

On the other hand, Atiyah and Witten [At] have found a remarkable formal link between the Index Theorem for Dirac operators on the spin complex and localization formulas in equivariant cohomology of Duistermaat-Heckman [DH], Berline-Vergne [BV1]. In particular the \hat{A} genus was interpreted in [At] as the inverse of an equivariant Euler form associated with an infinite-dimensional bundle. This suggested that an alternative approach to the Index Theorem, in relation with the equivariant cohomology of the loop space, was possible.

In [B2], we verified that the Atiyah-Witten formalism could be extended to the case of general Dirac operators, and also to fixed point theory. Also we showed in [B4] that the heat equation method is by itself such a reasonable proof of these formulas in infinite dimensions that it has a finite-dimensional counterpart, i.e., there is a proof of the formulas of [BV1] and [DH] which is at each step the finite-dimensional analogue of the probabilistic proof of the Index Theorem. This proof exhibits Patodi-like cancellations in finite dimensions. Conversely, it clearly demonstrates the purely geometric nature of these cancellations in Index Theory, the geometry to be considered being the geometry of the loop space.

Until recently, the heat equation formula (0.1) for the Index was considered a tool, which happened to work. The introduction of superconnections by Quillen [Q1] changed the situation dramatically. In [Q1], Quillen introduced a new class of objects, the superconnections on Z_2 graded finite-dimensional bundles, which makes (0.1) cry out to be considered as a formula for a Chern character. To briefly explain the analogy, let us just say that if E is a bundle with connection ∇ , if ∇^2 is the curvature of E , then $\text{ch } E$ is represented in cohomology by

$$\text{ch } E = \text{Tr}[\exp(-\nabla^2/2i\pi)]. \quad (0.2)$$

In [B3], we gave heat equation proofs of the Index Theorem of Atiyah-Singer for families of Dirac operators [AS], based on an infinite-dimensional analogue of Quillen's theory. To find the right choice of a superconnection, the finite-dimensional baby model of [B4] was of critical importance.

In relation with papers by Quillen [Q2] and Witten [W3] on determinant bundles and global anomalies, a transgressed form of Quillen's superconnection formalism has been introduced in Bismut-Freed [BF]. In particular, a remarkable argument of Witten [W3] relating the holonomy of determinant bundles to eta invariants has received a complete proof in [BF]. Another proof has recently been given by Cheeger [Ch].

Superconnections and the local form of the Index Theorem for families are currently used by Gillet and Soulé [GS] to construct direct images in Arakelov theory.

On the other hand, the results of Witten [W2] on the Morse inequalities, and the asymptotic Morse inequalities of Demailly for complex manifolds [De] have

also been proved by us [B6, B7] using heat equation methods. Getzler [Ge3, Ge4] has given a degree-theoretic interpretation in infinite dimensions of certain Index problems. Current efforts are done to relate in a more direct way heat equation methods to the cyclic homology of Connes [Co].

This paper is organized in the following way. In §1, the current heat equation proofs of the Index Theorem for Dirac operators are briefly reviewed. The new proofs have been classified into

- proofs related to supersymmetry,
- probabilistic proofs,
- group-theoretic proofs.

The principle of the probabilistic proof is briefly described, to emphasize its relations with the localization formulas in equivariant cohomology of Duistermaat-Heckman [DH], Berline-Vergne [BV1]. These relations are made explicit in §2, along the lines of Atiyah [At] and ourselves [B2, B4].

In §3, we briefly describe Quillen's superconnections [Q1] and their applications to the heat equation proof of the Atiyah-Singer Index Theorem for families of Dirac operators [B3]. One application to anomalies is also briefly indicated [BF].

I. The heat equation proofs of the Index Theorem. In this section, we briefly review the heat equation proofs of the Index Theorem for Dirac operators.

In (a) and (b), we summarize the now-classical proofs which rely on algebraic arguments.

In (c), we indicate some of the ideas involved in the recent proofs in [Ge1, Ge2, B1, BV2].

(a) *The heat equation method.* Let M be a compact connected Riemannian manifold of even dimension $n = 2l$. Let $E = E_+ \oplus E_-$ be a Z_2 graded complex Hermitian bundle over M , such that E_+ and E_- are orthogonal. Let τ be the involution of E defining the grading, i.e., $\tau = \pm 1$ on E_{\pm} .

$\Gamma(E)$, $\Gamma(E_{\pm})$ denote the sets of C^{∞} sections of E , E_{\pm} . Clearly $\Gamma(E) = \Gamma(E_+) \oplus \Gamma(E_-)$ is also naturally Z_2 graded. We still denote by τ the involution defining the grading in $\Gamma(E)$.

Also $\Gamma(E)$ can be endowed with the L_2 Hermitian product

$$h, h' \rightarrow \int_M \langle h, h' \rangle(x) dx. \quad (1.1)$$

Let D_+ be a first-order elliptic differential operator mapping $\Gamma(E_+)$ into $\Gamma(E_-)$. Let D_- be the formal adjoint of D_+ . Set

$$D = \begin{bmatrix} 0 & D_- \\ D_+ & 0 \end{bmatrix}. \quad (1.2)$$

End $\Gamma(E)$ is naturally Z_2 graded, the even (resp. odd) elements commuting (resp. anticommuting) with τ .

Clearly

$$D^2 = \begin{bmatrix} D_- D_+ & 0 \\ 0 & D_+ D_- \end{bmatrix}. \quad (1.3)$$

For any $t > 0$, $\exp(-tD^2/2)$ is given by a C^∞ kernel $P_t(x, y)$, so that if $h \in \Gamma(E)$

$$\exp\left(-\frac{tD^2}{2}\right) h(x) = \int_M P_t(x, y) h(y) dy. \quad (1.4)$$

For any $x \in M$, $P_t(x, x)$ is even in $\text{End}_x E$.

If A is a trace class operator acting on $\Gamma(E)$, we define its supertrace $\text{Tr}_s[A]$ by

$$\text{Tr}_s[A] = \text{Tr}[\tau A].$$

Recall that the index $\text{Ind } D_+$ of D_+ is given by

$$\text{Ind } D_+ = \dim \ker D_+ - \dim \ker D_-. \quad (1.5)$$

The first step in the calculation of $\text{Ind } D_+$ is the McKean-Singer formula [MS, ABP]:

$$\text{Ind } D_+ = \text{Tr}_s \left[\exp\left(-\frac{tD^2}{2}\right) \right] = \int_M \text{Tr}_s[P_t(x, x)] dx. \quad (1.6)$$

Let $P_t^\pm(x, x)$ be the restriction of $P_t(x, x)$ to $E_{\pm, x}$. Well-known results on zêta functions [Se] and heat kernels [ABP] show that as $t \downarrow 0$, for any $k \in \mathbb{N}$,

$$\text{Tr}[P_t^\pm(x, x)] = \sum_{j=-n/2}^k a_j^\pm(x) t^j + o(t^k, x). \quad (1.7)$$

In (1.6), $(a_j^\pm(x))$ are C^∞ functions which only depend on the local symbol of D . For $j \geq -n/2$, set

$$a_j(x) = a_j^+(x) - a_j^-(x). \quad (1.8)$$

Clearly

$$\text{Tr}_s[P_t(x, x)] = \sum_{j=-n/2}^k a_j(x) t^j + o(t^k, x). \quad (1.9)$$

From (1.5), (1.6), we find [MS, ABP],

$$\int_M a_j(x) dx = 0, \quad j \neq 0. \quad (1.10)$$

$$\text{Ind } D_+ = \int_M a_0(x) dx.$$

McKean and Singer [MS] conjectured that if $D = d + d^*$ acting on the de Rham complex, some extraordinary cancellations would show that for $j < 0$, $a_j = 0$, and that a_0 is exactly equal to the Chern-Gauss-Bonnet integrand for the Euler characteristic. In [P1], Patodi showed that this was indeed the case. In [P2], he extended his results to the Riemann-Roch theorem for Kähler manifolds.

(b) *Gilkey's theory of invariants.* In [Gi1], Gilkey established an algebraic theory of invariants. He showed that if $D = d + d^*$, the functions (a_j) belong to a certain class of local functions of the metric. After classifying such functions, Gilkey proved on a priori grounds that for $j < 0$, $a_j = 0$. The identification of a_0 was done in [Gi1] in an indirect way. Also Gilkey [Gi1] extended his approach to the Hirzebruch signature theorem.

In [ABP], Atiyah, Bott, and Patodi systematized the arguments of Gilkey to obtain the same type of result for twisted signature complexes, and derived the general Index Theorem. They developed Gilkey's theory in the realm of Riemannian geometry.

This point of view is systematically described in Gilkey's recent book [Gi3]. The theory of invariants has been also successfully applied [Gi2, 3] to prove the Lefschetz fixed point formulas of Atiyah-Bott [AB1] and Atiyah-Singer [AS].

(c) *Direct proofs of the cancellations and identification of the local integrand.* We now assume that M is orientable and spin. $F = F_+ \oplus F_-$ denotes the Z_2 graded Hermitian bundle of spinors over M . The Levi-Civita connection ∇^L of TM lifts into a unitary connection on F .

Let ξ be a complex Hermitian bundle over M , endowed with a unitary connection ∇^ξ .

Set $E = F \otimes \xi$, $E_\pm = F_\pm \otimes \xi$. E_\pm are Hermitian bundles, naturally endowed with the connection $\nabla^L \otimes 1 + 1 \otimes \nabla^\xi$, which we denote by ∇ .

Recall that if $e \in TM$, e acts on F by Clifford multiplication. $E = E_+ \oplus E_-$ is then a TM Clifford module.

We now define the Dirac operator. Let e_1, \dots, e_n be an orthonormal base of TM .

DEFINITION 1.1. D denotes the operator acting on $\Gamma(E)$,

$$D = \sum_1^n e_i \nabla_{e_i}. \quad (1.11)$$

D_\pm is the restriction of D to $\Gamma(E_\pm)$. D_\pm maps $\Gamma(E_\pm)$ into $\Gamma(E_\mp)$.

Let R be the curvature of TM , K the scalar curvature of M , L the curvature of ξ . Let Δ^H be the horizontal Laplacian on $\Gamma(E)$. Lichnerowicz's formula [Li] asserts that

$$D^2 = -\Delta^H + \frac{K}{4} + \frac{1}{2} \sum e_i e_j \otimes L(e_i, e_j). \quad (1.12)$$

1. *The supersymmetric proofs.* We first briefly review the arguments of Witten [W], Alvarez-Gaumé [Al], Friedan-Winney [FW], and Zumino [Z] leading to a supersymmetric derivation of the Index Theorem for D_+ .

Let LM be the loop space of M . The idea is to rewrite (1.6) in the form

$$\text{Ind } D_+ = \int_{LM} \exp\{\mathcal{L}^t(x)\} dD(x), \quad (1.13)$$

where $\mathcal{L}^t(x)$ is a supersymmetric Lagrangian, and $dD(x)$ is the “volume element” of LM . Let us just say that $\mathcal{L}^t(x)$ is a Lagrangian involving anticommuting variables $\psi, \bar{\psi}$. Supersymmetry here means that \mathcal{L}^t is invariant under transformations which involve x and the anticommuting variables $\psi, \bar{\psi}$. By making $t \downarrow 0$, and using arguments in particular from spectral theory, [A1, FW, Z] derive the local formula for the index

$$\text{Ind } D_+ = \int_M \hat{A}\left(\frac{R}{2\pi}\right) \text{Tr} \left[\exp - \frac{L}{2i\pi} \right]. \tag{1.14}$$

In (1.14) \hat{A} is the Hirzebruch polynomial on antisymmetric matrices:

$$\hat{A}(C) = \prod_1^l \frac{x_i/2}{\sinh(x_i/2)}. \tag{1.15}$$

In [Ge1], Getzler gave a rigorous formulation to the previous arguments. He used the supermanifold T^*M to give an asymptotic representation of the supertrace $\text{Tr}_s[P_t(x, x)]$ in terms of the graded symbol of $\exp(-tD^2/2)$. The local formula for the index is finally obtained by using a quadratic Gaussian approximation.

Recently, Getzler [Ge2] has given a new proof of the local convergence closely related to [Ge1] and also to [B1]. In [Ge2], Getzler adequately rescales the time, space, and Clifford variables to show that if D^ε is adequately rescaled with the factor ε , $(D^\varepsilon)^2$ converges to the partial differential operator on $T_{x_0}M$

$$\mathcal{L} = - \sum_1^n \left(\nabla_{e_i} + \frac{1}{4} R_{x_0}(e_i, x_j e_j) \right)^2 + L.$$

The explicit computations of the fundamental solution of $\partial/\partial t + \mathcal{L}$ leads again to the formula (1.14).

2. *The probabilistic proof.* We now briefly summarize our proof of the Index Theorem [B1]. To simplify, we assume that ξ is here the trivial bundle C .

Let $p_t(x, y)$ be the scalar heat kernel on M . Let E_{x_0, x_0}^t be the law on $C([0, 1]; M)$ of the Brownian bridge x_s^t starting at x_0 at time 0 and ending at x_0 at time 1 associated with the scaled metric g_M/t [B5, §2].

Let $\tau_0^{1,t}$ be the parallel transport operator from F_{x_0} into F_{x_0} along the loop x_s^t . An easy application of Itô’s formula [B1] shows that

$$\text{Tr}_s[P_t(x_0, x_0)] = p_t(x_0, x_0) E_{x_0, x_0}^t \left[\exp \left\{ - \frac{t \int_0^1 K(x_s^t) ds}{8} \right\} \text{Tr}_s[\tau_0^{1,t}] \right]. \tag{1.16}$$

As $t \downarrow 0$,

$$p_t(x_0, x_0) \simeq 1/(\sqrt{2\pi t})^n. \tag{1.17}$$

Also using the techniques of [B5], we describe E_{x_0, x_0}^t by means of a Brownian bridge w_s^1 in $T_{x_0}M$ with $w_0^1 = w_1^1 = 0$, so that approximately

$$x_s^t \sim \exp_{x_0} \left(\sqrt{t} w_s^1 \right). \tag{1.18}$$

Then $\tau_0^{1,t}$ acting on $T_{x_0}M$ has the expansion

$$\tau_0^{1,t} = I - \frac{t}{2} \int_0^1 R_{x_0}(dw^1, w^1) + o(t). \quad (1.19)$$

An argument from representation theory shows that

$$\frac{\text{Tr}_s \tau_0^{1,t}}{t^{n/2}} \rightarrow (-i)^t \text{Pf} \left[\frac{\int_0^1 R_{x_0}(dw^1, w^1)}{2} \right]. \quad (1.20)$$

We thus find that

$$\lim_{t \downarrow 0} \text{Tr}_s [P_t(x_0, x_0)] = \int \text{Pf} \left[\frac{-i}{4\pi} \int_0^1 R_{x_0}(dw^1, w^1) \right] dP_1(w^1). \quad (1.21)$$

If η is the Riemannian orientation form of TM , we get

$$\lim_{t \downarrow 0} \text{Tr}_s [P_t(x_0, x_0)] \eta(x_0) = \int \exp^\wedge \left\{ \frac{-i}{4\pi} \int_0^1 R_{x_0}(dw^1, w^1) \right\} dP_1(w^1), \quad (1.22)$$

where $\exp^\wedge \{ \dots \}$ is the exponential in $\Lambda(T^*M)$ of the corresponding 2 form. Using well-known symmetries of R , we find that

$$\lim_{t \downarrow 0} \text{Tr}_s [P_t(x_0, x_0)] \eta(x_0) = \int \exp^\wedge \left\{ \frac{-i}{4\pi} \int_0^1 \langle R_{x_0}(\cdot, \cdot) w^1, dw^1 \rangle \right\} dP_1(w^1). \quad (1.23)$$

A formula of P. Lévy [Le], known as the stochastic area formula, shows that the r.h.s. of (1.23) is equal to $\hat{A}(R/2\pi)$.

The Lefschetz fixed point formulas of Atiyah-Bott [AB1] and Atiyah-Singer [AS] were also proved in [B1] using the same sort of arguments and formulas of P. Lévy [Le].

3. *The group-theoretic proof.* In [BV], Berline and Vergne have given a proof of the Index formula and of the Lefschetz formulas by considering the scalar heat kernel on the bundle of orthonormal frames of TM . This idea is of course motivated by the $G \rightarrow G/H$ situation in group theory. The \hat{A} polynomial appears naturally in [BV2], being related to the jacobian of the exponential mapping in $\text{SO}(n)$.

II. Index Theorem and equivariant cohomology of the loop space.

In this section, we discuss the relations of the Index Theorem for Dirac operators to the equivariant cohomology of the loop space.

In (a), we summarize the observations of Atiyah and Witten [At]. In (b), we describe the baby model of [B4], where a proof of the localization formulas of [BV1, DH] is given, which is strictly parallel to the proof of [B1]. Patodi's cancellations in finite dimensions are exhibited.

(a) *The remark of Atiyah and Witten.* We now summarize the observation in [At].

Namely, the space LM of smooth loops $s \in R/Z \rightarrow x_s \in M$ is an infinite-dimensional manifold with the Riemannian metric

$$Y \in T_x LM \rightarrow \int_0^1 |Y_s|^2 ds.$$

S_1 acts naturally on LM by $x \rightarrow k_t x_0 = x_{\cdot+t}$, and the k_t are isometries. Let X be the Killing vector field generating k , so that

$$X(x)_s = dx/ds. \quad (2.1)$$

Let X' be the 1 form on LM ;

$$Y \in TLM \rightarrow X'(Y) = \langle X, Y \rangle. \quad (2.2)$$

One easily verifies that if $Y \in TLM$

$$dX'(Y, Z) = 2 \int_0^1 \left\langle \frac{DY}{Ds}, Z \right\rangle ds, \quad (2.3)$$

where DY/Ds is the covariant derivative of Y along x for the Levi-Civita connection.

The parallel transport operator τ_0^1 along the loop x acts like an element of $SO(n)$ on $T_{x_0}M$. Let $\pm\theta_j$ be the angles of τ_0^1 . One verifies easily that the eigenvalues of D/Ds acting on $T_x LM$ are given by

$$\pm 2i\pi m \pm i\theta_j, \quad m \in N. \quad (2.4)$$

The Pfaffian $\text{Pf}(-dX'/2)$ is given formally by

$$\text{Pf}\left(-\frac{dX'}{2}\right) = \prod_{j=1}^l \theta_j \prod_1^{+\infty} [4\pi^2 m^2 - \theta_j]^2. \quad (2.5)$$

Dividing (2.5) formally by the infinite $(\prod_1^l 4\pi^2 m^2)^l$, we get

$$\frac{\text{Pf}(-dX'/2)}{\prod_1^{+\infty} (4\pi^2 m^2)^l} = \prod_1^l 2\sin\left(\frac{\theta_j}{2}\right). \quad (2.6)$$

On the other hand, a formula from representation theory shows that if $\text{Tr}_s[\tau_0^1]$ is the supertrace of τ_0^1 acting on F_{\pm, x_0} ,

$$\text{Tr}_s[\tau_0^1] = \pm(i)^l \prod_1^l 2\sin\frac{\theta_j}{2}. \quad (2.7)$$

Using (1.6), (2.6), and (2.7), Atiyah gives the following formal formula

$$\text{Ind } D_+ = \frac{(\prod_1^{+\infty} m^2)^l}{(2\pi)^l} i^l \int_{LM} \exp\left\{-\frac{(d+i_X)X'}{2t}\right\}. \quad (2.8)$$

Also $L_X X' = 0$, and so

$$(d+i_X)[(d+i_X)X'] = 0. \quad (2.9)$$

The differential form μ_t appearing in the r.h.s. of (2.9) is such that $(d+i_X)\mu_t = 0$.

Now observe that $M = \{X = 0\}$.

Assume temporarily that LM is instead a finite-dimensional compact manifold. A formula of Duistermaat-Heckman [DH] and Berline-Vergne [BV1] (also

see [AB2]) asserts that if e is the equivariant Euler class of the normal bundle to M in LM , then

$$\int_{LM} \mu_t = \int_M \frac{\mu_t}{e}. \quad (2.10)$$

Now by calculating e formally in terms of the Levi-Civita connections on M and LM , one shows easily that $\hat{A}(R/2\pi)$ represents in cohomology

$$\frac{\prod_1^{+\infty} (m^2)^l}{(2\pi)^l} i^l \frac{1}{e}.$$

Also $\mu_t = 1$ on M_0 . We thus find that, rather surprisingly, a formal application of the formula of [DH, BV1] on LM “proves” the Index Theorem for the Dirac operators on the spin complex.

In [B2], we have shown that the observations of [At] extend to the case of Dirac operators acting on twisted spin complexes.

(b) *From infinite to finite dimensions: Patodi’s cancellations in finite dimensions.* We noticed in [B4] that formula (2.8) could lead to a proof of the localization formulas of [BV1, DH] which would be strictly parallel to the heat equation proof of the Index Theorem.

In fact let N be a compact Riemannian orientable manifold. X is a Killing vector field, X' the corresponding 1 form. N^X is the submanifold $N^X = (X = 0)$. μ is a smooth section of $\Lambda(T^*N)$ such that $(d + i_X)\mu = 0$. We claim that for any $s \geq 0$

$$\int_N \mu = \int_N \exp\{-s(d + i_X)X'\} \mu. \quad (2.11)$$

In fact the derivative of the r.h.s. of (2.1) is given by

$$-\int_M (d + i_X)[X' \wedge \exp\{-s(d + i_X)X'\} \mu] = 0, \quad (2.12)$$

and so for any $t > 0$

$$\int_N \mu = \int_N \exp\left\{-\frac{(d + i_X)X'}{2t}\right\} \mu. \quad (2.13)$$

As $t \downarrow 0$, the integral in the r.h.s. localizes on N^X . To make the analogy with §1c, we now assume that $\mu = 1$. Then

$$\int_N \exp - \frac{(d + i_X)X'}{2t} = \int_N \exp\left\{-\frac{|X|^2}{2t}\right\} \frac{\text{Pf}(-dX'/2)}{t^{\dim N/2}} dx. \quad (2.14)$$

If B is the normal bundle of N^X in N , let J^X be the infinitesimal action of X in B . By taking geodesic coordinates in the normal bundle and doing the change of variables $y = \sqrt{t}y'$ in B , we find that as $t \downarrow 0$, (2.14) is close to

$$\int_{N^X} dx \int_B \exp\left\{-\frac{|X|^2(x, y\sqrt{t})}{2t}\right\} \frac{\text{Pf}_{TN^X}[-dX'(x, y\sqrt{t})/2]}{t^{\dim N^X/2}} [\text{Pf}_B J_X] dy. \quad (2.15)$$

Let R be the curvature of TN . B is stable under $R(Y, Z)$, for $Y, Z \in TN^X$. Since X is Killing, $\nabla_Y(\nabla \cdot X) + R(X, Y) = 0$, and so

$$\frac{\text{Pf}_{TN^X}[-dX'(x, \sqrt{t}y)/2]}{t^{\dim N^X/2}} \rightarrow \text{Pf}_{TN^X} \left[-\frac{R}{2}(J_X y, y) \right]. \quad (2.16)$$

So as $t \downarrow 0$, (2.14) converges—while staying constant—to

$$\int_{N^x} \int_B \exp^{\wedge} \left\{ -\frac{|J_X y|^2}{2} - \frac{R}{2}(J_X y, y) \right\} \text{Pf}_B[J_X] dy. \quad (2.17)$$

At this stage the similarity of (2.17) with (1.20) and (1.22) should be obvious. (2.16) is a version of Patodi's cancellations in finite dimensions. It also gives a geometric origin to such cancellations.

III. Superconnections and the families Index Theorem. We now describe Quillen's superconnections [Q1] and their applications to the Index Theorem for families [B3, BF].

In (a), we describe the results of Quillen [Q1]. In (b), we summarize our heat equation proof of the Atiyah-Singer Index Theorem for families of Dirac operators [B3]. In (c), we summarize the results of [BF], in relation with [Q1, W3].

(a) *Quillen's superconnections.* Let N be a connected manifold. $E = E_+ \oplus E_-$ is a Z_2 graded vector bundle on N . $\text{End } E \hat{\otimes} \Lambda(T^*N)$ is a Z_2 graded algebra. The supertrace Tr_s defined on $\text{End } E$ extends to $\text{End } E \hat{\otimes} \Lambda(T^*N)$ and takes its values in $\Lambda(T^*N)$. Let ∇ be a connection on E preserving the grading. ∇ defines a first-order differential operator acting on smooth sections of $\Lambda(T^*N) \hat{\otimes} E$.

Let u be an odd smooth section of $\text{End } E \hat{\otimes} \Lambda(T^*N)$. $\nabla + u$ is a superconnection in the sense of Quillen [Q1]. $(\nabla + u)^2$ is an even section of $\Lambda(T^*N) \hat{\otimes} \text{End } E$ and is the curvature of $\nabla + u$.

We now have the result of Quillen [Q1].

THEOREM 3.1. $\text{Tr}_s \exp\{-(\nabla + u)^2/2\}$ is a closed form on N which is a representative of the scaled Chern character of $E_+ - E_-$.

In particular if D is an odd section of $\text{End } E$, $\nabla + D$ is a superconnection.

In [Q1], Quillen used superconnections to study differential forms and K -theory with support conditions and was also motivated by the Index Theorem for families. Mathai and Quillen [MQ] have used superconnections to study various problems related to localization and Thom forms.

(b) *The heat equation proof of the Index Theorem for families of Dirac operators.* Formula (1.6) for $\text{Ind } D_+$ is now crying out to be considered as a formula for a Chern character in the special case of one single operator.

In fact let $M \xrightarrow{\pi} B$ be a fibering of compact manifolds, with compact connected fibers Z of even dimension $n = 2l$. We assume that TZ is spin. Let g_Z be a smooth metric on TZ .

Let $F = F_+ \oplus F_-$ be the bundle of spinors of TZ . Let ξ be a Hermitian bundle on M , endowed with a unitary connection ∇^ξ .

For each $y \in B$, there is a well-defined Dirac operator

$$D_y = \begin{bmatrix} 0 & D_{-,y} \\ D_{+,y} & 0 \end{bmatrix}$$

on Z_y .

The Atiyah-Singer Index Theorem for families [AS] calculates $\ker D_+ - \ker D_- \in K(B)$.

In [B3], we have adapted Quillen's formalism in an infinite-dimensional situation. For $y \in B$, let $H_y^\infty = H_{+,y}^\infty \oplus H_{-,y}^\infty$ be the Z_2 graded bundle of C^∞ sections of $F \otimes \xi$ over Z_y . D is odd in $\text{End } H^\infty$.

Let $T^H M$ be a subbundle of TM such that $TM = T^H M \oplus TZ$. $T^H M$ identifies with π^*TB . Any metric g_B on TB lifts to $T^H M$. Let ∇^L be the Levi-Civita connection on TM endowed with the metric $g_B \oplus g_Z$. If P_Z is the projection operator from TM on TZ , let ∇^Z be the Euclidean connection on TZ ,

$$\nabla^Z = P_Z \nabla^L. \quad (3.1)$$

We proved in [B3] that ∇^Z does not depend on g_B , and is canonically defined by $T^H M$ and g_Z . ∇^Z and ∇^ξ define a unitary connection ∇ on $F \otimes \xi$.

For $Y \in TB$, let Y^H be the lift of Y in $T^H M$. If $h \in H^\infty$, set

$$\tilde{\nabla}_Y h = \nabla_{Y^H} h. \quad (3.2)$$

$\tilde{\nabla}$ is a connection on H^∞ . For any $t > 0$, $\tilde{\nabla} + \sqrt{t}D$ is a superconnection on H^∞ . The curvature $(\tilde{\nabla} + \sqrt{t}D)^2$ is a second-order elliptic operator acting fiberwise.

The following result is proved in [B3].

THEOREM 3.2. *For any $t > 0$, $\text{Tr}_s[\exp\{-(\tilde{\nabla} + \sqrt{t}D)^2/2\}]$ is a C^∞ closed form on B , which represents the scaled Chern character of $\ker D_+ - \ker D_-$.*

As $t \downarrow 0$, $\text{Tr}_s[\exp\{-(\tilde{\nabla} + \sqrt{t}D)^2/2\}]$ does not converge in general.

Let S be defined by

$$\nabla^L - \nabla^B \oplus \nabla^Z = S.$$

Let e_1, \dots, e_n be an orthonormal base of TZ . f_1, \dots, f_m is a base of TB which lifts into a base of $T^H M$; dy^1, \dots, dy^M is the corresponding dual base. In [B3, §3], we introduce the Levi-Civita superconnection

$$\begin{aligned} \tilde{\nabla}^{L,t} + \sqrt{t}D = \sum_{i,j,\alpha,\beta} \left[e_i \left(\sqrt{t} \nabla_{e_i} + \frac{1}{2} \langle S(e_i) e_j, f_\alpha \rangle e_j dy^\alpha \right. \right. \\ \left. \left. + \frac{1}{4\sqrt{t}} \langle S(e_i) f_\alpha, f_\beta \rangle dy^\alpha dy^\beta \right) \right. \\ \left. + dy^\alpha \left(\nabla_{f_\alpha} + \frac{1}{2\sqrt{t}} \langle S(f_\alpha) e_i, f_\beta \rangle e_i dy^\beta \right) \right]. \end{aligned} \quad (3.3)$$

Let K be the scalar curvature of the fiber Z and L the curvature of ξ .

The following formula is proved in [B3, §3].

THEOREM 3.3. *The curvature of the Levi-Civita superconnection is given by*

$$\begin{aligned}
 (\tilde{\nabla}^{L,t} + \sqrt{t}D)^2 = & -t \left(\nabla_{e_i} + \frac{1}{2t} \langle S(e_i)e_j, f_\alpha \rangle \sqrt{t}e_j dy^\alpha \right. \\
 & \left. + \frac{1}{4t} \langle S(e_i)f_\alpha, f_\beta \rangle dy^\alpha dy^\beta \right)^2 \\
 & + \frac{tK}{4} + \frac{1}{2} te_i e_j \otimes L(e_i, e_j) + \frac{1}{2} dy^\alpha dy^\beta \otimes L(f_\alpha, f_\beta) \\
 & + \sqrt{t}e_i dy^\alpha \otimes L(e_i, f_\alpha). \tag{3.4}
 \end{aligned}$$

We prove in [B3, §4] that as $t \downarrow 0$, $\text{Tr}_s[\exp\{-(\tilde{\nabla}^{L,t} + \sqrt{t}D)^2/2\}]$ converges. More precisely, we obtain a local version of this convergence. After adequately scaling the limit, we find that if R^Z is the curvature of TZ , the rescaled limit is

$$\int_Z \hat{A}\left(\frac{R^Z}{2\pi}\right) \text{Tr} \exp\left[-\frac{L}{2i\pi}\right] \tag{3.5}$$

We thus find that (3.5) represents $\text{ch}(\ker D_+ - \ker D_-)$. Recently, Berline and Vergne [BV3] have given a different proof of the convergence, using group-theoretic ideas.

(c) *Determinant bundles and the holonomy theorem.* In [Q2] Quillen has constructed a metric and a holomorphic connection on the determinant bundle of a family of $\bar{\partial}$ operators on Riemann surfaces. This construction has been extended in Bismut-Freed [BF] to the case of the family of Dirac operators considered in §3(b).

Namely, set

$$\lambda = (\det \ker D_+)^* \otimes \det \ker D_-. \tag{3.6}$$

λ is a well-defined C^∞ line bundle on B , even if B is noncompact [Q2, BF].

The first result of [BF] is that if the bundle λ is endowed with the Quillen metric, there is a unitary connection ${}^1\nabla$ on λ , whose curvature is given by

$$2i\pi \left[\int_Z \hat{A}\left(\frac{R^Z}{2\pi}\right) \text{Tr} \exp - \frac{L}{2i\pi} \right]^{(2)}. \tag{3.7}$$

In [W3] Witten has given an argument showing that in certain situations, the holonomy of a loop c in B could be calculated using the η -invariant of a Dirac operator on the cylinder $\pi^{-1}(c)$.

This result has been fully proved in [BF] for a family of Dirac operators. Recall that the η -function of a selfadjoint elliptic operator has been defined in Atiyah-Patodi-Singer [APS].

THEOREM 3.4. *Let c be a smooth loop in B . For $\varepsilon > 0$, let D'^ε be the Dirac operator on $\pi^{-1}(c)$ associated with the metric $g_B/\varepsilon \oplus g_Z$. Let $\eta^\varepsilon(s)$ be the η -function of D'^ε . Set*

$$\bar{\eta}^\varepsilon = (\eta^\varepsilon(0) + \dim \ker D'^\varepsilon)/2.$$

Then as $\varepsilon \downarrow 0$, $[\bar{\eta}^\varepsilon]$ has a limit $[\bar{\eta}]$ in R/Z . Also if τ is the holonomy of λ over c for the connection ${}^1\nabla$, then

$$\tau = (-1)^{\text{Ind } D_+} \exp(-2i\pi[\bar{\eta}]). \tag{3.8}$$

The result of Theorem 3.4 is strongly connected with Atiyah-Donnelly-Singer [ADS]. A new proof of Theorem 3.4 has been recently given by Cheeger [Ch].

REFERENCES

- [Al] L. Alvarez-Gaume, *Supersymmetry and the Atiyah-Singer Index Theorem*, Comm. Math. Phys. **90** (1983), 161–173.
- [At] M. F. Atiyah, *Circular symmetry and stationary phase approximation*, Proceedings of the conference in honor of L. Schwartz, Astérisque, no. 131, So. Math. France, Paris, 1985, pp. 43–59.
- [AB1] M. F. Atiyah and R. Bott, *A Lefschetz fixed point formula for elliptic complexes*. I, Ann. of Math. **86** (1967), 374–407; II, Ann. of Math. **88** (1968), 451–491.
- [AB2] —, *The moment map and equivariant cohomology*, Topology **23** (1984), 1–28.
- [ABP] M. F. Atiyah, R. Bott, and V. K. Patodi, *On the heat equation and the Index Theorem*, Invent. Math. **19** (1973), 279–330.
- [ADS] M. F. Atiyah, H. Donnelly, and I. M. Singer, *Eta invariants, signature defect of cusps and values of L functions*, Ann. of Math. **118** (1983), 131–177.
- [APS] M. F. Atiyah, V. K. Patodi, and I. M. Singer, *Spectral asymmetry and Riemannian geometry*. I, Math. Proc. Cambridge Philos. Soc. **77** (1975), 43–69.
- [AS] M. F. Atiyah and I. M. Singer, *The Index of elliptic operators*. I, Ann. of Math. **87** (1968), 484–530; III, **87** (1968), 546–604; IV, **93** (1971), 119–138.
- [B1] J. M. Bismut, *The Atiyah-Singer theorems: a probabilistic approach*. I, J. Funct. Anal. **57** (1984), 56–99; II, **57** (1984), 329–348.
- [B2] —, *Index Theorem and equivariant cohomology on the loop space*, Comm. Math. Phys. **98** (1985), 213–237.
- [B3] —, *The Index Theorem for families of Dirac operators: two heat equation proofs*, Invent. Math. **83** (1986), 91–151.
- [B4] —, *Localization formulas, superconnections and the Index Theorem for families*, Comm. Math. Phys. **103** (1986), 127–166.
- [B5] —, *Large deviations and the Malliavin calculus*, Progress in Math., no. 45, Birkhäuser, Boston, 1984.
- [B6] —, *The Witten complex and the degenerate Morse inequalities*, J. Differential Geom. **23** (1986), 207–240.
- [B7] —, *On Demailly's asymptotic Morse inequalities: a heat equation proof*, J. Funct. Anal. (to appear).
- [BF] J. M. Bismut and D. S. Freed, *The analysis of elliptic families*. I, Comm. Math. Phys. **106** (1986), 159–176; II, **107** (1986), 103–163.
- [BV1] N. Berline and M. Vergne, *Zéros d'un champ de vecteurs et classes caractéristiques équivariantes*, Duke Math. J. **50** (1983), 539–549.
- [BV2] —, *A computation of the equivariant index of the Dirac operator*, Bull. Soc. Math. France **113** (1985), 305–345.
- [BV3] —, *A proof of Bismut local index theorem for a family of Dirac operators*, Preprint IHES/M/86/46.
- [Ch] J. Cheeger, *Eta invariants, the adiabatic approximation and conical singularities*, (to appear).
- [Co] A. Connes, *Noncommutative differential geometry*, Inst. Hautes Études Sci. Publ. Math. **62** (1985), 41–144.
- [De] J. P. Demailly, *Champs magnétiques et inégalités de Morse pour la d' cohomologie*, Ann. Inst. Fourier (Grenoble) **35** (1985), 189–229.
- [DH] J. J. Duistermaat and G. Heckman, *On the variation in the cohomology of the symplectic form of the reduced phase space*, Invent. Math. **69** (1982), 259–268; Addendum **72** (1983), 153–158.
- [FW] D. Friedan and H. Windey, *Supersymmetric derivation of the Atiyah-Singer Index and the chiral anomaly*, Nuclear Phys. B **235** (1984), 395–416.
- [GS] H. Gillet and C. Soulé (to appear).

- [Ge1] E. Getzler, *Pseudodifferential operators on supermanifolds and the Atiyah-Singer Index Theorem*, Comm. Math. Phys. **92** (1983), 163–178.
- [Ge2] —, *A short proof of the Atiyah-Singer Index Theorem*, Topology **25** (1986), 111–117.
- [Ge3] —, *Degree Theory for Wiener maps*, J. Funct. Anal. **68** (1986), 388–403.
- [Ge4] —, *The degree of the Nicolai map in supersymmetric quantum mechanics*, J. Funct. Anal. (to appear).
- [G11] P. Gilkey, *Curvature and the eigenvalues of the Laplacian*, Adv. in Math. **10** (1973), 344–382.
- [G12] —, *Lefschetz fixed point formulas and the heat equation*, Partial Differential Equations and Geometry (Proc. Conf., Park City, Utah, 1977), C. Byrnes, editor, Lecture Notes in and Pure Appl. Math., vol. 48, Dekker, New York, 1979, pp. 91–147.
- [G13] —, *Invariance theory, the heat equation and the Atiyah-Singer Index Theorem*, Publish or Perish, Washington, 1984.
- [Le] P. Lévy, *Wiener's random functions, and other Laplacian random functions*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (J. Neyman, editor), Univ. of California Press, Berkeley, Calif., 1951, pp. 171–187.
- [Li] A. Lichnerowicz, *Spineurs harmoniques*, C. R. Acad. Sci. Paris Sér A **257** (1963), 7–9.
- [MQ] V. Mathai and D. Quillen, *Superconnections, Thom classes, and equivariant differential forms*, Topology **25** (1986), 85–110.
- [MS] H. McKean and I. M. Singer, *Curvature and the eigenvalues of the Laplacian*, J. Differential Geom. **1** (1967), 43–69.
- [P1] V. K. Patodi, *Curvature and the eigenforms of the Laplacian*, J. Differential Geom. **5** (1971), 233–249.
- [P2] —, *Analytic proof of the Riemann-Roch-Hirzebruch theorem for Kaehler manifolds*, J. Differential Geom. **5** (1971), 251–283.
- [P3] —, *Holomorphic Lefschetz fixed point formulas*, Bull. Amer. Math. Soc. **79** (1973), 825–828.
- [Q1] D. Quillen, *Superconnections and the Chern character*, Topology **24** (1985), 89–95.
- [Q2] —, *Determinants of Cauchy-Riemann operators over a Riemann surface*, Funct. Anal. Appl. **19** (1985), 31–34.
- [Se] R. T. Seeley, *Complex powers of an elliptic operator*, Proc. Sympos. Pure Math., vol. 10, Amer. Math. Soc., Providence, R.I., 1967, pp. 288–307.
- [W1] E. Witten, unpublished.
- [W2] —, *Supersymmetry and Morse theory*, J. Differential Geom. **17** (1982), 661–692.
- [W3] —, *Global gravitational anomalies*, Comm. Math. Phys. **100** (1985), 197–229.
- [Z] B. Zumino, *Supersymmetry and the Index Theorem*, Shelter Island II Conference (R. Jackiw et. al. editors), M. I. T. Press, Cambridge, Mass., 1985, pp. 79–94.

UNIVERSITÉ DE PARIS-SUD, MATHÉMATIQUES-BÂTIMENT 425, 91405 ORSAY CÉDEX, FRANCE

A Survey of Riemannian Metrics with Special Holonomy Groups

ROBERT L. BRYANT

0. Holonomy. We let M^n denote a smooth, connected, simply connected n -manifold and let g be a smooth Riemannian metric on M . For every $x \in M$, we let $H_x \subseteq \mathrm{SO}(T_x M)$ denote the holonomy of g at x . Here $\mathrm{SO}(T_x M)$ is the group of rotations of $T_x M$, and H_x is a closed, connected Lie subgroup of $\mathrm{SO}(T_x M)$. Parallel translation along any C^1 path from x to y induces an isomorphism $H_x \simeq H_y$. It follows that, if we choose an isometry $\iota: T_x M \rightarrow \mathbf{R}^n$, then the induced group $H = \{\iota \circ h \circ \iota^{-1} \mid h \in H_x\} \subseteq \mathrm{SO}(n)$ is uniquely defined up to conjugacy (by $\mathrm{O}(n)$) in $\mathrm{SO}(n)$. By abuse of language, we say that the holonomy of g is H , the representation of H into $\mathrm{SO}(n)$ being understood.

Because of the deRham splitting theorem (see [KN]), we are mainly interested in the cases where H acts irreducibly on \mathbf{R}^n . In 1955, M. Berger [Be] gave a list of all the possible irreducible holonomy groups for Riemannian metrics. We may state his theorem as follows:

THEOREM (BERGER). *Let (M^n, g) denote a connected, simply connected Riemannian manifold and suppose that for some $x \in M$, H_x acts irreducibly on $T_x M$. Then either g is a locally symmetric metric or else H is conjugate (in $\mathrm{O}(n)$) to one of the following standard subgroups of $\mathrm{SO}(n)$:*

- (i) $\mathrm{SO}(n)$,
- (ii) $\mathrm{U}(m)$ if $n = 2m > 2$,
- (iii) $\mathrm{SU}(m)$ if $n = 2m > 2$,
- (iv) $\mathrm{Sp}(m)\mathrm{Sp}(1)$ if $n = 4m > 4$,
- (v) $\mathrm{Sp}(m)$ if $n = 4m > 4$,
- (vi) G_2 if $n = 7$,
- (vii) $\mathrm{Spin}(7)$ if $n = 8$,
- (viii) $\mathrm{Spin}(9)$ if $n = 16$.

Later, Simons [Si] gave a different proof of this result in the Riemannian case. (Berger's original proof also covered the "metrics" of other signatures.)

Now the condition that g have holonomy H is a rather complicated integro-differential set of conditions. In practice, one usually works with the associated

H -structure. For each $y \in M$, let γ be a C^∞ path from y to x and let $P_\gamma: T_y \rightarrow T_x$ be g -parallel translation. Letting $\iota: T_x M \rightarrow \mathbf{R}^n$ be as before, we set $B_y = \{\iota \circ P_\gamma \mid \gamma \text{ is a } C^\infty \text{ path from } y \text{ to } x\}$. Then $B = \bigcup_y B_y$ is a smooth H -structure on M (we let $h \in H$ act on the *right* by letting $R_h(b) = h^{-1} \circ b$ for $b \in B$). The bundle B is the smallest subbundle of the g -orthonormal frame bundle which is invariant under g -parallel translation. In particular, the Levi-Civita connection of g restricts as a compatible, torsion-free connection on the H -structure B . Conversely, let $\mathcal{F} \rightarrow M$ denote the general frame bundle. Thus an element $u \in \mathcal{F}_x$ is an isomorphism $u: T_x M \rightarrow \mathbf{R}^n$ and \mathcal{F} is a right $\mathrm{GL}(n)$ bundle. If $H \subseteq \mathrm{GL}(n)$ is a closed subgroup, we say that an H -structure $B \subseteq \mathcal{F}$ is *torsion-free* if it possesses a compatible, torsion-free connection. For our purposes, it suffices to take $H \subseteq \mathrm{SO}(n)$ so we assume this henceforth. In this case, the only possible compatible, torsion-free connection on such a B is the restriction of the underlying Levi-Civita connection. The H -structures on M are the sections of the bundle $\mathcal{F}/H \rightarrow M$. It is easy to see [Bry2] that the torsion-free condition is a set of $n \dim(\mathrm{SO}(n)/H)$ first-order partial differential equations for the sections of $\mathcal{F}/H \rightarrow M$. Since an H -structure is torsion-free iff it is invariant under the Levi-Civita parallelism of the underlying metric, we see that the study of metrics with holonomy a subgroup of H (an “integro-differential” problem) is equivalent to the study of torsion-free H -structures (a “differential” problem). This simplifies matters considerably. Moreover, in each of the cases above, the underlying H -structure is interesting in its own right. An alternative adjective to “torsion-free” is “1-flat” since “torsion-free” is equivalent to the condition that the H -structure be “flat to first-order.” For details, see [Bry2].

Finally, we mention the *holonomy principle*. If (M, g) is a Riemannian manifold and $\iota: T_x M \rightarrow \mathbf{R}^n$ is an isometry inducing the group $H \subseteq \mathrm{SO}(n)$ corresponding to the holonomy group $H_x \subseteq \mathrm{SO}(T_x M)$, then any tensor on \mathbf{R}^n which is H -invariant gives rise to a parallel tensor field on M by first transferring by ι^{-1} to $T_x M$ and then parallel translating. For example, if ϕ_0 is an H -invariant p -form on \mathbf{R}^n , then ϕ_0 gives rise to a parallel (and hence harmonic) p -form $\phi \in \Omega^p(M)$. If H leaves fixed a complex structure $j: \mathbf{R}^n \rightarrow \mathbf{R}^n$ (n even), then j gives rise to a parallel (and hence integrable) almost complex structure $J: TM \rightarrow TM$. These remarks will be quite useful in the sequel. For more details, see [Bour].

1. Metrics with holonomy $\mathrm{SO}(n)$. Consider the curvature tensor as a mapping $R: \Lambda^2(TM) \rightarrow \mathrm{End}(TM)$. It follows from the Ambrose-Singer holonomy theorem (see [KN]) that, for every $x \in M$, we have $\mathrm{im}(R_x) \subseteq \mathfrak{h}_x \subseteq \mathfrak{so}(T_x M)$ where \mathfrak{h}_x is the Lie algebra of $H_x \subseteq \mathrm{SO}(T_x M)$. Since $\dim(\Lambda^2(T_x M)) = \dim(\mathfrak{so}(T_x M)) = n(n-1)/2$, it easily follows that whenever R_x is injective we must have $H_x = \mathrm{SO}(T_x M)$, i.e., $H = \mathrm{SO}(n)$. The injectivity of R_x is obviously an open condition on R_x and depends only on the 2-jet of the metric g at x . It follows easily that the space of metrics on M with holonomy $H = \mathrm{SO}(n)$ is both open and dense in the space of metrics on M . This is the meaning of the common statement that “the generic metric on M has holonomy $\mathrm{SO}(n)$.” Some

manifolds M^n have the property that *every* metric on M has holonomy $\mathrm{SO}(n)$. For example, S^n cannot be written as a product and hence H_x always acts irreducibly on $T_x S^n$. Moreover, each of the groups other than $\mathrm{SO}(n)$ on Berger's list have fixed some p -form for $0 < p < n$. This would give rise to nontrivial cohomology in $H^p(S^n, \mathbf{R})$ if there were a metric with $H \subsetneq \mathrm{SO}(n)$ on S^n . Since $H^p(S^n) = 0$ unless $p = 0$ or n , this is impossible.

2. Metrics on M^{2m} with holonomy $\mathrm{U}(m)$. The group $\mathrm{U}(m) \subseteq \mathrm{SO}(2m)$ is the set of complex linear orthogonal transformations of $\mathbf{C}^m = \mathbf{R}^{2m}$ with the Hermitian inner product $\langle z, w \rangle = \bar{z}^1 w^1 + \bar{z}^2 w^2 + \cdots + \bar{z}^m w^m$. Writing $\langle z, w \rangle = z \cdot w + \omega(z, w)i$, where $z \cdot w, \omega(z, w) \in \mathbf{R}$, we see that $\mathrm{U}(m)$ can also be defined as the subgroup of $\mathrm{O}(2m)$ which fixes the 2-form ω . Of course, ω and the complex structure are equivalent in the presence of the inner product. For example, $\omega(x, y) = (ix) \cdot y$.

By the holonomy principle, if (M^{2m}, g) has holonomy $H \subseteq \mathrm{U}(m)$, then M has a parallel orthogonal complex structure $J: TM \rightarrow TM$. Conversely, if (M, g) possesses a parallel complex structure, then $H \subseteq \mathrm{U}(m)$. Such a metric is called *Kähler*. In general, a $\mathrm{U}(m)$ -structure on M^{2m} is specified by an almost complex structure $J: TM \rightarrow TM$ and a nondegenerate 2-form $\omega \in \Omega^2(M)$ satisfying $\omega(Jx, y) = \omega(Jy, x)$ and the positivity condition $\omega(x, Jx) > 0$. In this case, the metric is given by $g(x, y) = \omega(Jx, y)$. The conditions that the $\mathrm{U}(m)$ -structure be torsion-free are simply that J be integrable and that ω be closed. For a more detailed analysis of these equations, see [Gray]. Suffice it to say here that these constitute $2m^2(m-1)$ equations of first order for the $3m^2$ unknowns which locally determine a section of the appropriate $\mathrm{GL}(2m)/\mathrm{U}(m)$ -bundle over M .

Of course, it is possible to integrate these equations explicitly. Since an integrable almost complex structure J is locally equivalent to the standard structure on \mathbf{C}^m , we may reduce our problem to finding the appropriate 2-form. It is well known that with respect to the standard complex coordinates z^1, \dots, z^m on \mathbf{C}^m , a closed $(1, 1)$ form ω can be written in the form (for some $f \in C^\infty(\mathbf{C}^m)$)

$$\omega = \sqrt{-1}(\partial^2 f / \partial z^a \partial \bar{z}^b) dz^a \wedge d\bar{z}^b.$$

If we set $H_f = (\partial^2 f / \partial z^a \partial \bar{z}^b) = {}^t \bar{H}_f$ (the Hermitian Hessian of f), then ω satisfies the positivity condition iff H_f is positive definite. Thus, after we reduce "change of coordinates" to "holomorphic change of coordinates," the metrics with holonomy $\mathrm{U}(m)$ "depend" on one function of $2m$ variables.

Of course, the literature on Kähler metrics is vast. We cannot attempt a recounting here, but refer the reader to standard texts such as [We] or [Gr-H].

3. Metrics on M^{2m} with holonomy $\mathrm{SU}(m)$. The group $\mathrm{SU}(m) \subseteq \mathrm{U}(m)$ is the subgroup of complex volume-preserving linear maps of \mathbf{C}^m . In addition to preserving the metric and the 2-form (of type $(1, 1)$), $\mathrm{SU}(m)$ preserves a complex volume form (of type $(n, 0)$).

If (M^{2m}, g) has holonomy $H \subseteq \mathrm{SU}(m)$, then in addition to a parallel orthogonal complex structure $J: TM \rightarrow TM$, M possesses a parallel complex volume

form $\nu \in \Omega^{n,0}(M)$. In particular, the canonical bundle of (M, J) is trivial and its first Chern form vanishes. This implies that (M, g) is Ricci-flat and hence that the metric g is real analytic in, say, holomorphic coordinates [dT-K].

Locally, since a complex volume form can always be put in normal form by, say, choosing coordinates so that $\nu = dz^1 \wedge \cdots \wedge dz^m$, we need only solve for ω . It turns out that ω can always be written in the form

$$\omega = \sqrt{-1}(\partial^2 f / \partial z^a \partial \bar{z}^b) dz^a \wedge d\bar{z}^b,$$

where f satisfies the Monge-Ampere equation

$$\det \left(\frac{\partial f}{\partial z^a \partial \bar{z}^b} \right) = 1$$

(see [Sem]). This is an elliptic second-order equation for f when f is holomorphically convex (i.e., $H_f > 0$). In the analytic category, f is determined by its value and first normal derivative along any real hypersurface in \mathbb{C}^m . Thus, the solutions depend on 2 real analytic functions of $2m - 1$ (real) variables (the ambiguity in the choice of volume-preserving holomorphic coordinates depends only on functions of m (real) variables).

At the global level, the fundamental result is Yau's solution of the Calabi conjecture: Any compact complex manifold with trivial canonical bundle which possesses *some* Kähler metric also possesses a Kähler metric for which the nonzero holomorphic volume form is parallel, i.e., a metric with holonomy $H \subseteq \mathrm{SU}(m)$. Examples of such manifolds are the smooth algebraic hypersurfaces in \mathbb{CP}^{m+1} of degree $m + 2$. Again, the literature on this subject is so vast that we cannot begin a recounting here. A good introduction is [Sem].

4. Metrics on M^{4m} with holonomy $\mathrm{Sp}(m)\mathrm{Sp}(1)$. The group $\mathrm{Sp}(m)\mathrm{Sp}(1) \subseteq \mathrm{SO}(4m)$ is the quotient of $\mathrm{Sp}(m) \times \mathrm{Sp}(1)$ by the group $\mathbf{Z}_2 = \{(\pm I_m, \pm 1)\}$. The action of $\mathrm{Sp}(m)\mathrm{Sp}(1)$ on \mathbf{R}^{4m} can be described as follows: Regard \mathbf{R}^{4m} as \mathbf{H}^m = columns of quaternions of height m . $\mathrm{Sp}(m)$ is the group of quaternion $m \times m$ matrices which satisfy ${}^t \bar{A} A = I_m$. Note that $\mathrm{Sp}(1)$ is simply the unit quaternions. We let $\mathrm{Sp}(m) \times \mathrm{Sp}(1)$ act on \mathbf{H}^m by $(A, q) \cdot v = (Av)\bar{q} = A(v\bar{q})$. (Note that \mathbf{H}^m is a *right* \mathbf{H} -vector space.) Clearly the \mathbf{Z}_2 -subgroup acts trivially, so we get an imbedding $\mathrm{Sp}(m)\mathrm{Sp}(1) \subseteq \mathrm{SO}(4m)$. It is easily seen that $\mathrm{Sp}(m)\mathrm{Sp}(1)$ fixes a 4-form ψ on \mathbf{R}^{4m} which can be described as follows: Let $\omega_I, \omega_J, \omega_K \in \Lambda^2((\mathbf{R}^{4m})^*)$ denote the 2-forms corresponding to the orthogonal complex structures induced on $\mathbf{H}^m = \mathbf{R}^{4m}$ by right multiplication by $i, j, k \in \mathbf{H}$. Then set $\psi = \frac{1}{6}(\omega_I^2 + \omega_J^2 + \omega_K^2)$. It is not hard to see that ψ generates the ring of $\mathrm{Sp}(m)\mathrm{Sp}(1)$ invariant forms on \mathbf{R}^{4m} (simply consider the cohomology of the symmetric space $\mathbf{HP}^m = \mathrm{Sp}(m+1)/\mathrm{Sp}(m) \times \mathrm{Sp}(1)$). Henceforth, we shall always assume $m > 1$ since in the case $m = 1$, $\mathrm{Sp}(1)\mathrm{Sp}(1) = \mathrm{SO}(4)$ and we are back in case 1.

A metric g on M^{4m} has holonomy $H \subseteq \mathrm{Sp}(m)\mathrm{Sp}(1)$ iff M has a *parallel* 4-form $\Psi \in \Omega^4(M)$ which, at each point, satisfies $\Psi_x = L^*(\psi)$ for some isometry $L: T_x M \rightarrow \mathbf{H}^m$. A metric with this property is called *quaternionic* (or *quaternionic Kähler*). To specify a $\mathrm{Sp}(m)\mathrm{Sp}(1)$ structure on M^{4m} it suffices to specify

the metric g and the compatible 4-form Ψ . The $\mathrm{Sp}(m)\mathrm{Sp}(1)$ -structure is torsion free iff Ψ is g -parallel. This is a set of $12m(m-1)(2m+1)$ first-order equations for the $(2m-1)(7m+3)$ unknowns which determine a section of the appropriate $\mathrm{GL}(4m)/\mathrm{Sp}(m)\mathrm{Sp}(1)$ bundle over M^{4m} . These equations have not been analyzed in detail and the generality of their solutions is not known. Note that it is not enough to assume that Ψ is closed and co-closed in order to insure that it be parallel with respect to its compatible metric.

Nevertheless, many examples of (M^{4m}, g) with holonomy $H \subseteq \mathrm{Sp}(m)\mathrm{Sp}(1)$ are known. For the known symmetric examples, see [Sa1] and [Wo]. For a recent construction of some new orbifold solutions via reduction, see [Ga-Law]. As of this writing the only compact smooth (M^{4m}, g) known with holonomy $H = \mathrm{Sp}(m)\mathrm{Sp}(1)$ are symmetric (either HP^m or quotients of its noncompact dual).

5. Metrics on M^{4m} with holonomy $\mathrm{Sp}(m)$. The group $\mathrm{Sp}(m) \subseteq \mathrm{SO}(4m)$ may be defined several ways. If we endow $\mathbf{R}^{4m} = \mathbf{H}^m$ with the \mathbf{H} -valued inner product $\langle x, y \rangle = \bar{x}_1 y_1 + \bar{x}_2 y_2 + \cdots + \bar{x}_m y_m$, then $\mathrm{Sp}(m)$ is the set of \mathbf{H} -linear transformations of \mathbf{H}^m which preserve $\langle \cdot, \cdot \rangle$. (In our convention, \mathbf{H}^m is a *right* \mathbf{H} -vector space.) We may expand $\langle \cdot, \cdot \rangle$ in the form

$$\langle x, y \rangle = x \cdot y + \omega_I(x, y)i + \omega_J(x, y)j + \omega_K(x, y)k,$$

where $x \cdot y, \omega_I(x, y), \dots, \omega_K(x, y)$ are real and bilinear. Thus, $\mathrm{Sp}(m)$ preserves three 2-forms as well as the real-valued inner product $x \cdot y$. A slightly different point of view which distinguishes $i \in \mathbf{H}$ is to write

$$\langle x, y \rangle = x \cdot y + \omega_I(x, y)i + j(\omega_J(x, y) - i\omega_K(x, y)).$$

If we regard \mathbf{H}^m as \mathbf{C}^{2m} , we can see that $\mathrm{Sp}(m)$ is the set of *complex linear* transformations which preserve the $(1, 1)$ (Hermitian) form $x \cdot y + \omega_I(x, y)i$ and the $(2, 0)$ (complex symplectic) form $\omega_J(x, y) - i\omega_K(x, y)$. Both points of view are useful.

By the holonomy principle, if (M^{4m}, g) has holonomy $H \subseteq \mathrm{Sp}(m) \subseteq \mathrm{SO}(4m)$, then TM possesses a parallel orthogonal quaternionic multiplication $TM \times \mathbf{H} \rightarrow TM$. Conversely, if TM has such a parallel multiplication, then clearly $H \subseteq \mathrm{Sp}(m)$. Note that, since $\mathrm{Sp}(m) \subseteq \mathrm{SU}(2m)$, we also have a complex structure on M corresponding to the multiplication by $i \in \mathbf{H}$. In fact, this generalizes to a whole S^2 -parameter family of complex structures on M , one for each imaginary unit quaternion. Also, since $\mathrm{Sp}(m) \subseteq \mathrm{Sp}(m, \mathbf{C})$, we see that M^{4m} can also be regarded as a complex symplectic manifold. Metrics with holonomy in $\mathrm{Sp}(m)$ are called *hyper-Kähler* or sometimes *Hamiltonian Kähler* (because of the symplectic structure).

If one is given an $\mathrm{Sp}(m)$ -structure on M^{4m} , it comes equipped with a metric g and three 2-forms, denoted $\omega_I, \omega_J, \omega_K$. A necessary and sufficient condition that the $\mathrm{Sp}(m)$ -structure have a torsion-free connection is that the three 2-forms be closed: $d\omega_I = d\omega_J = d\omega_K = 0$. It is easy to calculate that this is $12m^2(2m-1)$ first-order equations for the $\mathrm{Sp}(m)$ -structure. Since $\mathrm{Sp}(m)$ -structures are given

by sections of a bundle over M with typical fiber $\mathrm{GL}(4m)/\mathrm{Sp}(m)$, we see that our conditions constitute $12m^2(2m-1)$ equations for $m(14m-1)$ unknowns. Even when $m=1$, this is overdetermined. However, the equations can be partially integrated as follows: For any solution of $d\omega_I = d\omega_J = d\omega_K = 0$, the form $\Omega = \omega_J - i\omega_K$ is a closed complex 2-form satisfying the conditions of the complex Darboux theorem. It then follows that there exist complex coordinates z^1, \dots, z^{2m} locally for which we may write

$$\Omega = dz^1 \wedge dz^{m+1} + dz^2 \wedge dz^{m+2} + \dots + dz^m \wedge dz^{2m}.$$

Since ω_I is, in particular, a Kähler-metric form on this complex manifold, it follows that we may write ω_I in the form $\omega_I = i(\partial^2 f / \partial z^a \partial \bar{z}^b) dz^a \wedge d\bar{z}^b$. The algebraic conditions on $\{\omega_I, \omega_J, \omega_K\}$ then force f to satisfy the conditions $\bar{H}_f J H_f = J$ and $H_f > 0$ where $H_f = (\partial^2 f / \partial z^a \partial \bar{z}^b)$ is the (Hermitian) Hessian of f and $J = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}$. It is easy to compute that this is a set of $m(2m-1)$ real equations on H_f and hence represents, for $m > 1$, an overdetermined system of second-order equations for the real function f . When $m=1$, this reduces to the complex Monge-Ampere equation. When $m > 1$, this overdetermined system is not involutive. In fact, these equations are now under investigation and will be reported on elsewhere.

According to [A1], the curvature tensor of a metric with holonomy $H \subseteq \mathrm{Sp}(m)$ takes values in a bundle of rank $\binom{2m+3}{4}$ modeled on an irreducible $\mathrm{Sp}(m)$ -module of dimension $\binom{2m+3}{4}$. Thus any such metric is Ricci-flat and hence real analytic in harmonic or holomorphic coordinates.

Apparently, Calabi [Ca] was the first to write down an explicit metric (on $T'^*\mathbf{CP}^m$) with holonomy $\mathrm{Sp}(m)$ for $m > 1$. This example is complete but not compact. Actually Bogomolov [Bo] had announced that there were no compact examples for $m > 1$. This turned out to be erroneous. Fujiki showed how one could construct a metric with holonomy $\mathrm{Sp}(2)$ on a complex 4-manifold got by resolving the singularities of the symmetric product of a $K3$ surface with itself. Later Beauville [Beau] showed how this could be generalized to the Douady powers $S^{[m]}$ for $m \geq 1$, where S is a $K3$ surface, thus yielding a metric on $M^{4m} = S^{[m]}$ with holonomy $\mathrm{Sp}(m)$ for all $m \geq 1$. It should be remarked that these constructions depend heavily on Yau's solution of the Calabi conjecture. For more details, we refer the reader to [Beau].

6. Metrics on M^7 with holonomy G_2 . Let e_1, \dots, e_7 be an oriented orthonormal basis of \mathbf{R}^7 . Let $\omega^1, \dots, \omega^7$ denote the dual 1-forms on \mathbf{R}^7 . We write ω^{ijk} as shorthand for $\omega^i \wedge \omega^j \wedge \omega^k$. Let $\phi \in \Lambda^3((\mathbf{R}^7)^*)$ be given by

$$\phi_0 = \omega^{123} + \omega^{145} + \omega^{167} + \omega^{246} - \omega^{257} - \omega^{356} - \omega^{347}.$$

Then by [Bry2], $G_2 \subseteq \mathrm{GL}(\mathbf{R}^7)$ is the subgroup which fixes ϕ_0 . Also $G_2 \subseteq \mathrm{SO}(7)$, it is simple, simply connected and of dimension 14. G_2 acts transitively on $S^6 \subseteq \mathbf{R}^7$; moreover, G_2 acts transitively on the Stiefel manifold of pairs of orthonormal vectors in \mathbf{R}^7 . G_2 leaves fixed ϕ_0 , $*\phi_0$, and the unit volume form

on \mathbf{R}^7 . It is also possible to define G_2 as the group of automorphisms of a vector cross product on \mathbf{R}^7 or as the group of automorphisms of the Cayley numbers (see [HL] or [Bry2]), but we will not need this.

If V is a vector space of dimension 7, we shall say that a form $\alpha \in \Lambda^3(V^*)$ is *positive* if there exists a linear isomorphism $L: V \rightarrow \mathbf{R}^7$ so that $\alpha = L^*(\phi_0)$. The set of positive forms $\Lambda_+^3(V^*) \subseteq \Lambda^3(V^*)$ is easily seen to be an open set (in $\Lambda^3(V^*)$) which is diffeomorphic to $GL(7)/G_2$.

By the holonomy principle, if (M^7, g) is a Riemannian manifold with holonomy $H \subseteq G_2$, then M^7 possesses a parallel 3-form, say $\phi \in \Omega^3(M)$, which is positive at every $x \in M$. Conversely, if (M^7, g) possesses a parallel positive 3-form, then (M^7, g) must have holonomy $H \subseteq G_2$. In [Bry2], it is shown that the holonomy H is a *proper* subgroup of G_2 iff (M^7, g) possesses a parallel 1-form.

The space of G_2 -structures on M^7 is clearly the same as the space of positive 3-forms on M^7 , $\Omega_+^3(M) =$ sections of the bundle $\Lambda_+^3(T^*M)$. This space is easily seen to be nonempty iff $w_1(M) = w_2(M) = 0$, i.e., if M is orientable and spin. Given a positive 3-form $\phi \in \Omega_+^3(M)$, we can define a metric and orientation on $T_x M$ for each x by requiring that a map $L: T_x M \rightarrow \mathbf{R}^7$ with $\phi_x = L^*(\phi_0)$ be an oriented isometry. Since $G_2 \subseteq SO(7)$, this is well defined. We denote the induced metric on M by g_ϕ and the induced Hodge star operator by $*_\phi$. Then it is known that the conditions $d\phi = d*_\phi \phi = 0$ are equivalent to the condition that ϕ be g_ϕ -parallel. In [Bry2], it is shown that the overdetermined system of equations $d\phi = d*_\phi \phi = 0$ for sections $\phi \in \Omega_+^3(M)$ is involutive, and a description of the local analytic solutions (via Cartan-Kähler theory) is given. Roughly, the germs of solutions to this system modulo the germs of local diffeomorphisms is “parametrized” by six functions of six variables. Moreover, for the “generic” solution ϕ , the holonomy H of g_ϕ is equal to G_2 .

As shown in [A1], the curvature tensor of a metric g with holonomy $H \subseteq G_2$ takes values in bundle of rank 77 modeled on an irreducible G_2 -module of dimension 77. It follows immediately that any such metric g is Ricci-flat and hence is real analytic in harmonic coordinates [dT-K]. The Cartan-Kähler analysis in [Bry2] shows that the curvature tensor at a point of M can assume any value in this bundle.

As of this writing, only one example of a metric with holonomy G_2 is known [Bry2]. This example is not complete, being the “cone metric” on the cone $\mathbf{R}^+ \times SU(3)/T^2$ over $SU(3)/T^2$ (T^2 is a maximal torus in $SU(3)$), where $SU(3)/T^2$ is given the appropriate bi-invariant metric. It is not known whether or not complete or compact examples exist.

7. Metrics on M^8 with holonomy $Spin(7)$. Keeping the notations of the previous section, we embed \mathbf{R}^7 into \mathbf{R}^8 and extend our basis by adding an e_0 . The dual 1-forms are, as before, $\omega^0, \dots, \omega^7$. Let $\phi_0 \in \Lambda^3((\mathbf{R}^7)^*)$ and $*\phi_0 \in \Lambda^4((\mathbf{R}^7)^*)$ be as defined before and set

$$\Phi = \omega^0 \wedge \phi_0 + (*\phi_0).$$

Then, by [Bry2], $\text{Spin}(7) \subseteq \text{GL}(8)$ is the subgroup which fixes Φ . It is known that $\text{Spin}(7) \subseteq \text{SO}(8)$, it is simply connected and is the double cover of the simple group $\text{SO}(7)$. $\text{Spin}(7)$ fixes only $\Phi (= *\Phi)$, 1, and $*1$ in $\Lambda^*((\mathbf{R}^8)^*)$. For other definitions of $\text{Spin}(7)$, see [HL] or [Fer].

If V is a vector space of dimension 8, we shall say that a form $\beta \in \Lambda^4(V^*)$ is *definite* if there exists a linear isomorphism

$$L: V \rightarrow \mathbf{R}^8$$

so that $\beta = L^*(\Phi)$. We denote the set of definite forms in $\Lambda^4(V^*)$ by $\Lambda_+^4(V^*)$. This latter space is diffeomorphic to $\text{GL}(8)/\text{Spin}(7)$.

By the holonomy principle, if (M^8, g) is a Riemannian manifold with holonomy $H \subseteq \text{Spin}(7)$, then M^8 possesses a parallel definite 4-form. The space of $\text{Spin}(7)$ -structures on M^8 is equivalent to the space of definite 4-forms $\Omega_+^4(M) =$ the sections of $\Lambda_+^4(T^*M)$. This space is nonempty iff M is orientable, spin, and the euler class of the positive spin bundle \mathbf{S}_+ is zero. Given a definite 4-form, say $\phi \in \Omega_+^4(M)$, there is a unique orientation and metric induced on M by requiring that $L: T_x M \rightarrow \mathbf{R}^8$ be an oriented isometry whenever $L^*(\Phi) = \phi_x$. We denote this metric by g_ϕ and the associated Hodge star by $*_\phi$. It is then known that the condition $d\phi = 0$ is equivalent to the condition that ϕ be g_ϕ -parallel. In [Bry2], it is shown that the overdetermined system of equations $d\phi = 0$ for sections $\phi \in \Omega_+^4(M)$ is involutive, and a description of the local (analytic) solutions (via Cartan-Kähler theory) is given. Roughly, the germs of solutions modulo the germs of local diffeomorphisms are “parametrized” by twelve functions of seven variables. Moreover, for the “generic” solution ϕ , the holonomy H of g_ϕ is equal to $\text{Spin}(7)$.

Again [A1] shows that the curvature tensor of a metric g with holonomy $H \subseteq \text{Spin}(7)$ takes values in a bundle of rank 168 modeled on an irreducible $\text{Spin}(7)$ -module of dimension 168. (In fact, this turns out to be the $\text{SO}(7)$ module of Weyl tensors of metrics on seven-manifolds.) It follows that any such metric g is Ricci-flat and hence is real analytic in harmonic coordinates. The Cartan-Kähler analysis in [Bry2] shows that the curvature tensor at a point of M can assume any value in this bundle.

As of this writing, there are two types of metrics with holonomy $\text{Spin}(7)$ known. The first is an isolated example, the “cone metric” on the cone $\mathbf{R}^+ \times \text{SO}(5)/\text{SO}(3)$ over $\text{SO}(5)/\text{SO}(3)$. ($\text{SO}(3)$ imbedded is a principal subgroup of $\text{SO}(5)$.) The second class of examples may be constructed on appropriate \mathbf{H}^* -bundles over local self-dual Einstein 4-manifolds with positive scalar curvature. (These later examples will be reported on elsewhere.) None of these examples is complete. It is not even known if compact or complete examples exist.

If compact examples exist, they are quite complicated. Aside from the obvious consequences for cohomology and fundamental group forced by Ricci-flatness, Lawson has observed that such a manifold would have $\hat{A} = 1$. This, coupled with the vanishing of the Euler class of \mathbf{S}_+ , forces severe restrictions on the

Pontrjagin numbers of such a manifold. This will be reported on in the above mentioned paper by Harvey and the author.

8. Metrics on M^{16} with holonomy $\text{Spin}(9)$. Let $\text{Spin}(9)$ denote the double cover of $\text{SO}(9)$. It is well known that $\text{Spin}(9)$ has a unique faithful irreducible representation of dimension 16. Thus, we may regard $\text{Spin}(9)$ as a subgroup of $\text{SO}(16)$ unique up to conjugacy in $\text{O}(16)$. Let $\mathfrak{so}(9) \subseteq \mathfrak{so}(16)$ denote the induced inclusion of Lie algebras. An elementary calculation, via roots and weights, shows that the space of curvature tensors $R: \Lambda^2(\mathbf{R}^{16}) \rightarrow \mathfrak{so}(16)$ which have values in $\mathfrak{so}(9)$ and which satisfy the Bianchi identities is of (real) dimension 1. (This calculation is due to Alexeevski [A1], but also, see [Br-Gr].) Let $R_0: \Lambda^2(\mathbf{R}^{16}) \rightarrow \mathfrak{so}(9) \subseteq \mathfrak{so}(16)$ be a nonzero element of this vector space. Note that $\text{OP}^2 = F_4/\text{Spin}(9)$ has holonomy $\text{Spin}(9)$ and is Einstein. Since its curvature tensor must be linearly equivalent to R_0 , it follows that R_0 represents an Einstein curvature tensor. It easily follows that any manifold (M^{16}, g) with holonomy $H \subseteq \text{Spin}(9)$ must be Einstein and have constant scalar curvature. Moreover, if the scalar curvature vanishes, then the entire curvature vanishes and hence (M^{16}, g) is flat. If (M^{16}, g) is not flat, then the curvature tensor at each point maps surjectively $R_x: \Lambda^2(T_x M) \rightarrow \mathfrak{so}(9)_x$ and hence we must have $H = \text{Spin}(9)$. Finally, since the holonomy algebra bundle, $\mathfrak{so}(9) \subseteq \Lambda^2(TM)$, is parallel, it easily follows that the curvature tensor is parallel. Thus *any metric g on M^{16} with holonomy $\text{Spin}(9)$ is symmetric*. Thus, the only examples are the constant multiples of the standard metrics on $F_4/\text{Spin}(9)$ and its noncompact dual. For more details consult [A1] or [Br-Gr].

BIBLIOGRAPHY

- [A1] D. V. Alexeevski, *Riemannian spaces with unusual holonomy groups*, Funktsional Anal. i Prilozhen **2** (1968), 1–10; English transl. Funct. Anal. Appl.
- [Beau] A. Beauville, *Varieties Kähleriennes dont la première classe de Chern est nulle*, J. Differential Geom. **18** (1983), 755–782.
- [Be] M. Berger, *Sur les groupes d'holonomie des variétés à connexion affine et des variétés Riemanniennes*, Bull. Soc. Math. France **83** (1955), 279–310.
- [Bo] F. A. Bogomolov, *Hamiltonian Kähler manifolds*, Dokl. Akad. Nauk SSSR **243** (1978), 1101–1104; English transl. in Soviet Math. Dokl. **19** (1978).
- [Bour] J. P. Bourguignon, *Groupes d'holonomie des variétés Riemanniennes*, Asterisque **126**, Soc. Math. France, 1985, pp. 169–180.
- [Br-Gr] R. B. Brown and A. Gray, *Riemannian manifolds with holonomy group $\text{Spin}(9)$* , Differential Geometry (in honor of K. Yano), Kinokuniya, Tokyo, 1972, 41–59.
- [Bry1] R. Bryant, *Metrics with holonomy G_2 or $\text{Spin}(7)$* , Lecture Notes in Math., vol. 1111, Springer-Verlag, Berlin, 1985, pp. 269–277.
- [Bry2] —, *Metrics with exceptional holonomy*, Preprint.
- [Ca] E. Calabi, *Isometric families of Kähler structures*, The Chern Symposium 1979 (Proc. Internat. Sympos., Berkeley, Calif., 1979), Springer-Verlag, 1980, pp. 23–39.
- [Ch] S. S. Chern, *On a generalization of Kähler geometry*, Algebraic Geometry and Topology: A Symposium in Honor of S. Lefschetz, Princeton Univ. Press, Princeton, N.J., 1957, pp. 103–121.
- [dT-K] D. De Turck and J. Kasdan, *Some regularity theorems in Riemannian geometry*, Ann. Sci. École Norm. Sup (4) **14** (1981), 249–260.

- [Fer] M. Fernandez, *A classification of Riemannian manifolds with structure group Spin(7)*, Preprint.
- [Fer-Gr] M. Fernandez and A. Gray, *Riemannian manifolds with structure group G_2* , Ann. Mat. Pura Appl. (4) **32** (1982), 19–45.
- [Ga-Law] K. Galicki and H. B. Lawson, *Quaternionic reduction and quaternionic orbifolds*, Preprint.
- [Gray] A. Gray and L. M. Hervella, *The sixteen classes of almost Hermitian manifolds and their linear invariants*, Ann. Mat. Pura Appl. (4) **123** (1980), 35–58.
- [Gr-H] P. Griffiths and J. Harris, *Principles of algebraic geometry*, Wiley-Interscience, 1978.
- [HL] R. Harvey and H. B. Lawson, *Calibrated geometries*, Acta Math. **148** (1982), 47–157.
- [KN] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*. Vols. 1 and 2, Wiley-Interscience, 1963.
- [Sal] S. Salamon, *Quaternionic Kähler manifolds*, Invent. Math. **67** (1982), 143.
- [Sem] *Première classe de Chern et courbure Ricci: preuve de la conjecture de Calabi* (Sem., Palaiseau, 1978), Asterisque, no. 58, Soc. Math. France, Paris, 1978.
- [SI] J. Simons, *On transitivity holonomy systems*, Ann. of Math. (2) **76** (1962), 213–234.
- [We] R. O. Wells, *Differential analysis on complex manifolds*, 2nd ed., Graduate Texts in Math., vol. 65, Springer-Verlag, 1980.
- [Wo] J. A. Wolf, *Complex homogeneous contact manifolds and quaternionic symmetric spaces*, J. Math. Mech. **14** (1965), 1033.

RICE UNIVERSITY, HOUSTON, TEXAS 77251, USA

On the Formulas of Atiyah-Patodi-Singer and Witten

JEFF CHEEGER

0. Introduction. The work of Atiyah-Patodi-Singer relates the η -invariants of certain natural elliptic differential operators with secondary invariants Chern-Simons type [APS, CS, ChS]. A recent formula of Witten calculates the η -invariants of such operators in a different way, for spaces which fibre over S^1 (see [W₁, W₂]). Here we explain how Witten's formula can be viewed as a consequence of

(i) a (renormalized) version of the Atiyah-Patodi-Singer formula for the operator $i\nabla_{\partial/\partial u}$ acting on sections of an infinite-dimensional Hermitian vector bundle over S^1 (where u is arclength on S^1),

(ii) the analog of a basic point in Quillen's generalization of the Weil homomorphism to superconnections [Q₂].

We also relate these ideas to analysis on spaces with conical singularities and (briefly) to secondary invariants of *volume collapses* with *bounded curvature*; see [CG₁, CG₂, CG₃, Y].

We are grateful to Ron Douglas and Blaine Lawson for a number of helpful comments.

1. The Atiyah-Patodi-Singer formula. In the present exposition we will consider only the following two cases of the Atiyah-Patodi-Singer formula. The discussion of the second case (and the corresponding formula of Witten) extends to other operators of Dirac type.

1.1. EXAMPLE. Let ∇ be a Hermitian connection on a Hermitian vector bundle F over the unit circle S^1 . If u denotes arclength on S^1 , then

$$\eta(i\nabla_{\partial/\partial u}) \equiv 2\hat{c}_1(\nabla)[S^1] \pmod{Z}. \quad (1.2)$$

Here, $\hat{c}_1(\nabla)[S^1]$ is the secondary geometric invariant associated to the first Chern class and the connection ∇ . It is just the phase θ , where $e^{2\pi i\theta}$ is the determinant of the holonomy of ∇ .

1.3. EXAMPLE. Let N^{4k-1} be a closed oriented riemannian manifold and let E be a Hermitian vector bundle with Hermitian connection, over N^{4k-1} . Let

Research supported by National Science Foundation Grant DMS-8405956.

$A: \Lambda^{\text{ev}}(N) \otimes E \rightarrow \Lambda^{\text{ev}}(N) \otimes E$ be defined by

$$A = (d * + (-1)^p * d) \quad \text{on } \Lambda^{2p}(N). \quad (1.4)$$

Suppose that $N^{4k-1} = \partial M^{4k}$, that E extends over M^{4k} , and that near ∂M^{4k-1} the metric on M^{4k} and connection on E split as products in the direction normal to ∂M^{4k-1} . Let $P_L(R)$, $P_{\text{ch}}(\Omega)$ denote the characteristic forms corresponding to the L -class of M^{4k} and Chern character of E . Then

$$\eta(A) \equiv \int_{M^{4k}} P_L(R) \Lambda P_{\text{ch}}(\Omega) \pmod{Z}. \quad (1.5)$$

The right-hand side is again an invariant of Chern-Simons type.

Neither invariant in (1.5) is canonically locally computable on N^{4k-1} . But if the data defining either side is varied, then the variational derivative is locally computable.

2. Riemannian submersions and the adiabatic limit. Let $Y \rightarrow X \xrightarrow{\pi} Z$ be a fibration of riemannian manifolds. Let g, h denote the metrics on X, Z respectively. Then π is called a *riemannian submersion* if

$$g = \pi^* h + k, \quad (2.1)$$

where k vanishes on the normal space to the fibers. Put

$$g_\delta = \pi^*(\delta^{-2}h) + k. \quad (2.2)$$

As $\delta \rightarrow 0$, the metrics $g_\delta, \delta^{-2}h$ blow up. Letting $\delta \rightarrow 0$ is called going to the *adiabatic limit*. On the other hand, in the terminology of [CG₃], the rescaled family $(X, \delta^2 g_\delta)$ is said to *collapse* to Z (where, in general, the curvature of $\delta^2 g_\delta$ blows up).

Let $\{\bar{\nabla}^\delta\}$ be the family of riemannian connections associated to $\{g_\delta\}$ (or equivalently, to $\{\delta^2 g_\delta\}$). Then it is easy to check the existence of

$$\bar{\nabla}^0 = \lim_{\delta \rightarrow 0} \bar{\nabla}^\delta. \quad (2.3)$$

Hence, if $X = N^{4k-1}$, the space of formula (1.5), the limit of the right-hand side of (1.5) as $\delta \rightarrow 0$ exists.

Under the additional assumption $Z = S^1$, Witten derived a remarkable formula for the limit of the left-hand side of (1.5). We describe this in §3. From the viewpoint of §3, passing to the adiabatic limit enters as a computational device (which, in fact, is not required for the approach described in §5). More importantly, however, it simplifies the formula (by killing an extra local term), thus making it appear more natural.

3. Witten's formula. Let N^{4k-1}, E be as in Example 1.3 and let $Y^{4k-2} \rightarrow N^{4k-1} \xrightarrow{\pi} S^1$ be a riemannian submersion. Let $\tilde{d}, \tilde{*}$ denote exterior differentiation (for E -valued forms) and the Hodge $*$ -operator, on a typical fibre, Y^{4k-2} . Put

$$\beta = \begin{cases} (-1)^p \tilde{*}, & \Lambda^{2p}(Y^{4k-2}), \\ \tilde{*}, & \Lambda^{2p-1}(Y^{4k-2}), \end{cases} \quad (3.1)$$

and

$$\mathcal{A} = \begin{cases} (-1)^{p+1}(\tilde{d}\beta + \beta\tilde{d}), & \Lambda^{2p}(Y^{4k-2}), \\ \tilde{d}\beta - \beta\tilde{d}, & \Lambda^{2p-1}(Y^{4k-2}). \end{cases} \quad (3.2)$$

Then \mathcal{A} is selfadjoint, elliptic, and

$$\beta^2 = -1, \quad (3.3)$$

$$\mathcal{A}\beta = -\beta\mathcal{A}. \quad (3.4)$$

Write

$$\begin{pmatrix} \phi \\ \omega \end{pmatrix} = \phi + du\Lambda\omega \in \Lambda^{2p}(X^{4k-1}), \quad (3.5)$$

where u is arclength on S^1 . Then the operator A of (1.4) is given by

$$A = \begin{pmatrix} \beta\partial_u & \mathcal{A} \\ \mathcal{A} & \partial_u\beta \end{pmatrix}. \quad (3.6)$$

Witten's formula for $\lim_{\delta \rightarrow 0} \eta(A_\delta)$ is locally computable on S^1 , if $\mathcal{A}(u)$ is invertible for all u . Put $\dot{\mathcal{A}} = \partial\mathcal{A}/\partial u$ and let $e^{-\mathcal{A}^2\varepsilon}$ denote the heat kernel of \mathcal{A}^2 . Then for the operator A_δ corresponding to g_δ ,

$$\lim_{\delta \rightarrow 0} \eta(A_\delta) \equiv \frac{1}{\pi} \int_{S^1} \lim_{\varepsilon \rightarrow 0} \text{tr} \left(\frac{\beta}{2} \mathcal{A}^{-1} \dot{\mathcal{A}} e^{-\mathcal{A}^2\varepsilon} \right) du. \quad (3.7)$$

The expression on the right-hand side occurs in physics in connection with the concept of *global anomaly* and with the related concept of *determinant line bundle* (see [AS, BF₁, BF₂, F, Q₁]). This was Witten's original motivation.

Witten's derivation of (3.7) used the adiabatic approximation in a form often employed in quantum mechanics. Rigorous treatments were given in [BF₂] using the heat equation and probability theory and in [C₄] using the heat equation and the finite propagation speed technique of [CT, CGT].

We now explain our interpretation of (3.1). We begin by regarding E -valued forms on N^{4k-1} as forms on S^1 with coefficients in the bundle Λ whose fibre at $u \in S^1$ is the (infinite-dimensional) space of forms on $\pi^{-1}(u)$. The metric on N^{4k-1} and connection on E induce a connection ∂ on Λ in the obvious way. Put $\alpha = \tilde{*}^{-1}\partial\tilde{*}$. Then

$$\alpha\beta = -\beta\alpha. \quad (3.8)$$

Moreover, if we set

$$\nabla_{\partial/\partial u} = \partial_u + \frac{1}{2}\alpha, \quad (3.9)$$

then ∇ is a Hermitian connection on Λ , for which β is parallel,

$$\nabla\beta \equiv 0. \quad (3.10)$$

Let Λ_\pm denote the $\pm i$ -eigenbundles of β on Λ . For the splitting $\Lambda = \Lambda_+ \oplus \Lambda_-$,

$$A_\delta = \begin{pmatrix} \delta i \nabla_{\partial/\partial u} & C_\delta \\ C_\delta & -\delta i \nabla_{\partial/\partial u} \end{pmatrix}, \quad (3.11)$$

where in terms of the splitting in (3.6),

$$C_\delta = \begin{pmatrix} -\delta\beta\frac{\alpha}{2} & \mathcal{A} \\ \mathcal{A} & \delta\beta\frac{\alpha}{2} \end{pmatrix}. \quad (3.12)$$

Since the objects above and relations (3.9)–(3.12) make equally good sense if $\dim \Lambda < \infty$, it should be possible to view (3.7) as the (renormalized) infinite-dimensional generalization of the corresponding formula in the finite-dimensional case.

Suppose $\dim \Lambda < \infty$. Then we can set $\varepsilon = 0$ in (3.7). Let $\delta\beta\nabla_{\partial/\partial u}$ denote the operator obtained by removing the off-diagonal part of A_δ . As usual, $\eta(\delta\beta\nabla_{\partial/\partial u}) = \eta(\beta\nabla_{\partial/\partial u})$ and we have

$$\beta\nabla_{\partial/\partial u} = \begin{pmatrix} i\nabla_{\partial/\partial u} & 0 \\ 0 & -i\nabla_{\partial/\partial u} \end{pmatrix}. \quad (3.13)$$

Applying (1.2) with $F = \Lambda_\pm$ gives

$$\eta(\beta\nabla_{\partial/\partial u}) \equiv 2\{\hat{c}_1(\nabla^+) - \hat{c}_1(\nabla^-)\}[S^1] \pmod{Z}, \quad (3.14)$$

where $\nabla^\pm = \nabla \mid \Lambda^\pm$. We claim that the right-hand side of (3.7) is just a way of explicitly rewriting (3.14). Granted (3.14) and this assertion, formula (3.7) is equivalent to

$$\lim_{\delta \rightarrow 0} \eta(A_\delta) \equiv \eta(\beta\nabla_{\partial/\partial u}) \pmod{Z}. \quad (3.15)$$

3.16. REMARK. The equality in (3.15) mirrors the generalization of the Weil homomorphism to *superconnections*, presented by Quillen. In fact, for $\dim \Lambda < \infty$ (but not, in general, for $\dim \Lambda = \infty$) one can verify that $\eta(A_\delta)$ is actually independent of S . Bismut and Freed emphasize superconnections and interpret the right-hand side of (3.7) in terms of the determinant line bundle. To reconcile this interpretation with that above, note that for a finite-dimensional Hermitian bundle (E^n, ∇)

$$\hat{c}_1(E^n, \nabla) = \hat{c}_1(\Lambda^n E, \nabla). \quad (3.17)$$

To see the relation between the right-hand sides of (3.7), (3.14), pull back ∇^- , via \mathcal{A} , to a connection $\mathcal{A}^*(\nabla^-)$ on Λ_+ . Then

$$\mathcal{A}^*(\nabla^-) = \mathcal{A}^{-1} \left(\frac{1+i\beta}{2} \right) \{(\nabla\mathcal{A}) + \mathcal{A}\alpha\} \quad (3.18)$$

and using (3.9),

$$\nabla^+ - \mathcal{A}^*(\nabla^-) = -\frac{1-i\beta}{2} \mathcal{A}^{-1} \nabla \mathcal{A}. \quad (3.19)$$

By the standard formula for the difference of Chern-Simons invariants,

$$\begin{aligned} 2[\hat{c}_1(\nabla^+) - \hat{c}_1(\nabla^-)] &= 2[\hat{c}_1(\nabla^+) - \hat{c}_1(\mathcal{A}^*(\nabla^-))] \\ &= \frac{-1}{\pi i} \int_{S^1} \text{tr} \left(\frac{1-i\beta}{2} \mathcal{A}^{-1} \nabla_{\partial/\partial u} \mathcal{A} \right) du. \end{aligned} \quad (3.20)$$

Using

$$\text{tr}(\mathcal{A}^{-1} \nabla_{\partial/\partial u} \mathcal{A}) = (\partial/\partial u) \log \det(\mathcal{A} \mid \Lambda), \quad (3.21)$$

and

$$\mathrm{tr}(\alpha\beta) = 0 \quad (3.22)$$

(which follows from (3.9)), the expression in (3.20) reduces to that in (3.7), if $\dim \Lambda < \infty$.

3.23. REMARK. If $\mathcal{A}: \Lambda_+ \rightarrow \Lambda_-$ is not invertible, (3.14) and (3.15) still hold. But then a (global) correction term coming from $\ker \mathcal{A}$, $\mathrm{coker} \mathcal{A}$ ($\mathcal{A}: \Lambda_+ \rightarrow \Lambda_-$) must be added to the right-hand side of (3.7).

3.24. REMARK. Considerations similar to the above apply in the case of higher-dimensional base spaces.

3.25. REMARK. The formula of [ADS] computes η -invariants associated to certain solve manifolds Σ in terms of related L -functions (see also [Mü]). The case of their result for 1-dimensional base spaces follows easily from (3.7) and Remark 3.23. Of course, some of the techniques used in deriving (3.7) are already present in [ADS]. The higher-dimensional case of the results of [ADS] will follow from the generalization of (3.7) alluded to in Remark 3.24.

4. An aside: collapse with bounded curvature. The family

$$(X^{4k-1}, \delta^2 g_\delta)$$

(see (1.27)) exhibits *volume collapse*. The limits of secondary geometric invariants for volume collapses with *bounded curvature* are studied in [CG₂, Y]. They are related to Bott residues, to polarized F -structures, and (along different lines) to von Neumann dimension. An F -structure is a generalization of an isometric abelian group action. For example, the fibrations $T^{4k-2} \rightarrow \Sigma^{4k-1} \rightarrow S^1$ with holonomy in $\mathrm{SL}(4k-2, Z)$ define natural polarized F -structures; compare Remark 3.25.

5. Conical singularities. Let

$$W^{4k-1} = N^{4k-1} \cup C_{0,1}(Y^{4k-2}) \quad (5.1)$$

be an oriented space with an isolated metrically conical singularity at $p \in X^{4k}$; see [C₁]. If

$$H^{2k-1}(Y^{4k-2}, R) = 0 \quad (5.2)$$

(which corresponds to the condition that \mathcal{A} of §1 is invertible), then the operator A of (1.4), on $W^{4k-1} \setminus p$, is essentially selfadjoint and its η -invariant is defined; see [C₃]. The right-hand side of (3.7) arose independently in this context (see [C₃, C₄]). Namely, let (W^{4k-1}, g_u) be a 1-parameter family of spaces as in (5.1). Let $\tilde{*}$ be the $*$ -operator on Y^{4k-2} and let P_{ce} denote orthogonal projection on coexact forms. Then

$$\frac{d}{du} \eta(W^{4k-1}, g_u) = \mathrm{loc}(X \setminus p) + \frac{1}{\pi} \int_{S^1} \lim_{\varepsilon \rightarrow 0} (\mathrm{tr} \tilde{*} P_{ce} e^{-\Delta_{2k-1} \varepsilon}). \quad (5.3)$$

Here, $\mathrm{loc}(W^{4k-1} \setminus p)$ is the usual local term and the second term is equal to the right-hand side of (3.7). (We are assuming that the bundle, E , is globally flat, although the discussion generalizes if this is not the case.)

A related interpretation is the following. Let X^{4k} be a space with nonisolated conical singularities and singular strata Σ^0, Σ^1 of dimension 0, 1. Let $N^{4k-1} \xrightarrow{\pi} S^1$ be the link of a component S^1 of Σ^1 . Suppose we compute the L_2 -signature of X^{4k} by the heat of equation method. Then the right-hand side of (3.7) is the contribution at Σ^1 to the constant term of the asymptotic expansion of the trace of the heat kernel (up to a correction term which is local on Σ').

5.4. REMARK. The generalization of (3.7) for higher-dimensional base spaces arises in connection with spaces having singular strata of higher dimension.

5.5. REMARK. The L_2 -signature of a space X^{4k} with conical singularities is equal to the signature in the sense of intersection homology; see [GM, C₂].

5.6. REMARK. Let W^{4k-1}, Y^{4k-2} be as in (5.1) and suppose that (5.3) does not hold. Then a choice of "ideal boundary condition" is required to define $\eta(A)$; see [C₁, C₂]. This choice also determines a selfadjoint transformation on the space of harmonic $(2k-1)$ forms, which anticommutes with β (as do A, C of (3.12)).

5.7. EXAMPLE. Let $N^{4k-1} \xrightarrow{\pi} S^1$ be the space of (3.7). Form a singular space X^{4k} , by starting with $N^{4k-1} \times [0, 1]$ and then attaching the cone on N^{4k-1} and the mapping cone of π to the boundary components $N \times 1, N \times 0$ respectively. Now (3.7) results from computing $\text{sig}_{L_2}(X^{4k})$ by the heat equation method.

5.8. EXAMPLE. Alternatively, (3.7) results from computing the L_2 -signature of $N^{4k-1} \times [0, 1]$, relative to two different global boundary conditions at $N \times 1, N \times 0$.

REFERENCES

- [AS] M. F. Atiyah and I. Singer, *Dirac operators coupled to vector potentials*, Proc. Nat. Acad. Sci. (1984), 2597.
- [ADS] M. F. Atiyah, H. Donnelly, and I. M. Singer, *Eta invariants, signature defects of cusps, and values of L-functions*, Ann. of Math. (2) **118** (1983), 131–177.
- [APS] M. F. Atiyah, V. K. Patodi, and I. M. Singer, *Spectral asymmetry and Riemannian geometry*. I, Math. Proc. Cambridge Philos. Soc. **77** (1975), 43–69; II, Math. Proc. Cambridge Philos. Soc. **78** (1975), 405–432; III, Math. Proc. Cambridge Philos. Soc. **79** (1976), 71–99.
- [BF₁] J. M. Bismut and D. S. Freed, *The analysis of elliptic families: Metrics and connections on determinant bundles*, Comm. Math. Phys. **106** (1986), 156–159.
- [BF₂] —, *The analysis of elliptic families: Dirac operators, eta invariants, and the holonomy theorem of Witten*, Comm. Math. Phys. **107** (1986), 103–163.
- [C₁] J. Cheeger, *On the spectral geometry of spaces with cone-like singularities*, Proc. Nat. Acad. Sci. U.S.A. **76** (1979), 2103–2106.
- [C₂] —, *On the Hodge theory of Riemannian pseudomanifolds*, Proc. Sympos. Pure Math., vol. 36, Amer. Math. Soc., Providence, R. I., 1980, pp. 91–145.
- [C₃] —, *Spectral geometry of singular Riemannian spaces*, J. Differential Geom. **18** (1983), 575–657.
- [C₄] —, *η -invariants, the adiabatic approximation and conical singularities*, J. Diff. Geom. (to appear).
- [CG₁] J. Cheeger and M. Gromov, *On the characteristic numbers of complete manifolds of bounded curvature and finite volume*, H. E. Ranch Memorial Volume, I. Chavel and H. Farkas, editors, Springer-Verlag, Heidelberg, 1985, pp. 115–154.
- [CG₂] —, *Bounds on the von Neumann dimension of L_2 -cohomology and the Gauss-Bonnet theorem for open manifolds*, J. Differential Geom. **20** (1984).

- [CG₃] —, *Collapsing Riemannian manifolds while keeping their curvature bounded*. I, J. Differential Geom. **23** (1986), 309–346.
- [CS] S. S. Chern and J. Simons, *Characteristic forms and geometric invariants*, Ann. of Math. (2) **99** (1974), 48–69.
- [ChS] J. Cheeger and J. Simons, *Differential characters and geometric invariants*, Geometry and Topology, Lecture Notes in Math., vol. 1167, Springer-Verlag, 1985.
- [CGT] J. Cheeger, M. Gromov, and M. Taylor, *Finite propagation speed, kernel estimates for functions of the Laplace operator, and the geometry of complete Riemannian manifolds*, J. Differential Geom. **17** (1982), 15–53.
- [CT] J. Cheeger and M. Taylor, *On the diffraction of waves by conical singularities*. I, Comm. Pure Appl. Math. **35** (1982), 275–331.
- [CT] —, *On the diffraction of waves by conical singularities*. II, Comm. Pure Appl. Math. **35** (1982), 487–529.
- [Chou] A. Chou, *The Dirac operator on spaces with conical singularities*, Ph.D. Thesis, SUNY, Stony Brook, 1982.
- [F] D. Freed, *Determinants, torsion and strings*, Preprint.
- [GM] M. Goresky and R. MacPherson, *Intersection homology theory*, Topology **19** (1980), 135–162.
- [Mü] W. Müller, *Signature defects of cusps of Hilbert modular varieties and values of L -series at $s = 1$* .
- [Q₁] D. Quillen, *Determinants of Cauchy-Riemann operators over a Riemann surface*, Funktsional. Anal. i Prilozhen. **19** (1985), 37.
- [Q₂] —, *Superconnections and the Chern character*, Topology **24** (1985), 89–95.
- [W₁] E. Witten, *Global gravitational anomalies*, Comm. Math. Phys. **100** (1985), 197–229.
- [W₂] —, *Global anomalies in string theory*, Symposium on Anomalies Geometry Topology, W. Bardeen and A. White, editors, World Scientific Press, 1985.
- [Y] D. G. Yang, *A residue theorem for secondary invariants of collapsed Riemannian manifolds*, Ph.D. Thesis, SUNY, Stony Brook, 1986.

STATE UNIVERSITY OF NEW YORK AT STONY BROOK, STONY BROOK, NEW YORK
11794, USA

Spectres de variétés riemanniennes et spectres de graphes

YVES COLIN DE VERDIÈRE

L'article de M. Kac "*Can one hear the shape of a drum?*" [KC] a été l'origine de nombreuses contributions au problème spectral inverse: déterminer certains invariants géométriques ou topologiques d'une variété riemannienne compacte à partir du spectre du laplacien. Moins d'attention a été portée au problème de savoir quelles sont les suites de valeurs propres possibles pour le laplacien d'une métrique riemannienne arbitraire sur une variété compacte donnée. D'autres problèmes du même type se posent avec la recherche d'ouverts bornés de \mathbf{R}^d tels que le laplacien euclidien avec conditions de Dirichlet ou de Neumann admette un spectre donné; ou encore avec l'opérateur de Schrödinger avec champ électrique et (ou magnétique) sur une variété riemannienne compacte donnée. Ces problèmes semblent d'accès très difficile comme le montrent les contraintes imposées par l'existence de différents développements asymptotiques (formules de Minakshisundaram-Pleijel, formule de Poisson). Nous nous restreindrons ici à l'étude du problème plus élémentaire de réaliser une suite finie

$$s_N = \{\lambda_1 = 0 < \lambda_2 \leq \dots \leq \lambda_N\}$$

comme suite des N premières valeurs propres avec multiplicité d'un opérateur à choisir dans une famille donnée d'opérateurs.

Indiquons quelques résultats typiques:

- Si X est une variété compacte de dimension ≥ 3 et s_N une suite arbitraire comme précédemment, il existe sur X des métriques riemanniennes telles que le laplacien admette cette suite comme suite des N premières valeurs propres avec multiplicité. En particulier, il n'y a aucune restriction sur les multiplicités des valeurs propres.

- Si X est une surface compacte, la situation est moins simple, car on sait [CG, BN] que la multiplicité de la première valeur propre non nulle du laplacien est majorée en fonction de la topologie de X . Les résultats obtenus montrent que cela semble être la seule contrainte; par exemple, toute suite $\{\lambda_1 = 0 < \lambda_2 < \dots < \lambda_N\}$ est la suite des N premières valeurs propres du laplacien pour une métrique bien choisie sur n'importe quelle surface.

• Pour ce qui est des laplaciens euclidiens dans les ouverts de \mathbf{R}^d , la situation est différente suivant le type de conditions aux limites:

Pour le problème de Dirichlet, les inégalités de Payne, Polya, et Weinberger [P-P-W] (et aussi [PR]) imposent des restrictions d'un autre type que la multiplicité: par exemple, pour tout ouvert borné de \mathbf{R}^d , les valeurs propres du problème de Dirichlet vérifient, pour tout n , $\lambda_{n+1} \leq 3\lambda_n$.

Pour le problème de Neumann, les seules restrictions sont de multiplicités: par exemple, pour toute suite

$$\{\lambda_1 = 0 < \lambda_2 < \dots < \lambda_N\},$$

il existe un ouvert lisse simplement connexe de \mathbf{R}^2 ayant cette suite comme suite des N premières valeurs propres du problème de Neumann.

Quelques indications sur les méthodes utilisées: comme les familles d'opérateurs étudiées ne sont pas normalisées (i.e. sont invariantes par homothétie), il suffit de trouver un opérateur ayant la suite

$$\varepsilon s_N = \{0 = \lambda_1 < \varepsilon \lambda_2 \leq \dots \leq \varepsilon \lambda_N\}$$

comme suite des N premières valeurs propres. On cherche à évaluer les petites valeurs propres d'un laplacien par *effet tunnel purement géométrique* (i.e. sans barrière de potentiel); cette évaluation met en jeu une *matrice dite d'interaction* qui peut s'interpréter géométriquement comme un *laplacien combinatoire* (avec poids) sur un graphe qui est le graphe d'adjacence d'un découpage de la variété en régions séparées par des passages étroits (tunnels!!). La matrice d'interaction est utilisée depuis longtemps en chimie pour l'étude de l'opérateur de Schrödinger gouvernant les électrons dans une molécule (méthode L.C.A.O.) et une étude mathématique très complète a été proposée récemment par Helffer-Sjöstrand [H-S].

Cette évaluation par perturbation d'un laplacien combinatoire ne permettrait pas a priori de réaliser exactement le spectre voulu: en effet, dans le cas d'une valeur propre multiple, par exemple, la perturbation entraîne une dispersion du paquet de valeurs propres, c'est ici qu'intervient une idée fondamentale: utiliser la transversalité et travailler uniformément avec des hamiltoniens dépendant d'un paramètre variant dans une boule de \mathbf{R}^p : V. Arnold a montré [AD] que dans une famille de matrices symétriques à coefficients réels l'apparition de valeurs propres multiples peut être stable. Pour mettre en oeuvre cette idée, on introduit une notion de *stabilité* d'une famille finie de valeurs propres pour un hamiltonien d'une famille à paramètres. Tous les spectres construits auront cette propriété, en particulier les exemples trouvés pourront avoir un caractère générique (pas d'isométries, etc.) et on pourra lisser les exemples non lisses.

Dans cette présentation, nous nous restreindrons à quelques points essentiels: énoncés précis des résultats, définitions relatives à la stabilité et à la matrice d'interaction, laplaciens combinatoires et exemples, étude de l'effet tunnel géométrique en dimension 2, méthodes de perturbations singulières.

1. Énoncés des résultats. Pour énoncer les résultats, il sera pratique d'introduire quelques notations: soit \mathcal{F} une famille (ensemble) d'opérateurs autoadjoints ≥ 0 , à résolvante compacte et ayant 0 comme valeur propre simple. Nous introduisons les propriétés suivantes de \mathcal{F} :

DÉFINITIONS. (i) On dira que \mathcal{F} vérifie (A_N) si, pour toute suite $\lambda_1 = 0 < \lambda_2 \leq \dots \leq \lambda_N$, il existe H dans \mathcal{F} ayant cette suite comme suite des N premières valeurs propres.

(ii) On dira que \mathcal{F} vérifie (B_N) si on a le même énoncé pour toute suite $\lambda_1 = 0 < \lambda_2 \dots < \lambda_N$.

(iii) On dira que \mathcal{F} vérifie (C_N) si pour toute suite $\mu_1 = 0 \leq \mu_2 \leq \dots \leq \mu_N$, il existe H dans \mathcal{F} et $b \geq 0$ tels que $\mu_1 + b \leq \mu_2 + b \leq \dots \leq \mu_N + b$ soit un intervalle du spectre de H .

Il est clair que la propriété (A_N) implique (B_N) et (C_N) . Si X est une variété compacte, nous noterons \mathcal{F}_X la famille des laplaciens associés à une métrique riemannienne C^∞ sur X . Pour $d \geq 2$, nous noterons \mathcal{G}_d la famille des laplaciens avec conditions de Neumann sur un ouvert borné, C^1 par morceaux de \mathbf{R}^d .

Nous avons alors les

THÉORÈME 1. Si X est de dimension ≥ 3 , \mathcal{F}_X vérifie (A_N) pour tout N .

En général posons $N(X) = \sup\{N | \mathcal{F}_X \text{ vérifie } (A_N)\}$.

THÉORÈME 2. Si X est de dimension 2, \mathcal{F}_X vérifie (B_N) et (C_N) pour tout N . On a $N(S^2) = 4$, $N(T^2) = 7$, $N(\mathbf{R}P^2) = 6$, $N(K^2) = 6$ et si X_g est la surface orientable de genre g , $g \geq 2$,

$$\max\left(7, E\left(3/2 + \sqrt{2g + 1/4}\right)\right) \leq N(X_g) \leq 4g + 4.$$

THÉORÈME 3. \mathcal{G}_d vérifie (A_N) pour toute N si $d \geq 3$. \mathcal{G}_2 vérifie (A_N) si et seulement si $N \leq 4$. \mathcal{G}_2 vérifie (B_N) pour tout N .

REMARQUE. Soit, pour une variété compacte X (avec ou sans bord), $C(X)$ le nombre chromatique de X défini comme le plus grand entier N tels qu'il existe un plongement du graphe complet à N sommets dans X ; alors dans tous les cas où l'on connaît $N(X)$ on a $N(X) = C(X)$ et les inégalités connues sont compatible avec les

CONJECTURE 1. Pour toute variété compacte X , on a $N(X) = C(X)$.

CONJECTURE 2. Pour toute variété compacte X , la multiplicité maximale de la première valeur propre non nulle du laplacien d'une métrique riemannienne sur X est $C(X) - 1$.

On rappelle [RL] que $C(X_g) = E(7/2 + \sqrt{12g + 1/4})$.

REMARQUE. Dans le cas de la bouteille de Klein K^2 la multiplicité maximale de λ_2 pour une métrique plate est 3. La multiplicité 5 est obtenue pour une métrique qui fait de K^2 un espace métrique proche de $\mathbf{R}P^2$ avec sa métrique usuelle (pour la distance de Hausdorff).

2. Stabilité des valeurs propres; matrice d'interaction. Introduisons d'abord une *notation*: si q est une forme quadratique de domaine D dans un espace de Hilbert \mathcal{H} et si E_0 et E sont 2 sous-espaces vectoriels de même dimension finie N de D , on notera q_E la restriction de q à E , et si E est assez proche de E_0 pour être le graphe d'une application B de E_0 dans E_0^\perp , on pose, pour $x \in E_0$, $\beta x = x \oplus Bx$, $B \in \mathcal{L}(E_0, E)$ et q_{E/E_0} la forme quadratique sur E_0 transportée de q_E par l'isométrie U de E_0 sur E définie par $U = \beta \circ (\beta^* \beta)^{-1/2}$. L'important est que q_{E/E_0} a sur $(E_0, \langle \cdot, \cdot \rangle)$ les mêmes valeurs propres que q_E sur $(E, \langle \cdot, \cdot \rangle)$.

Soit maintenant $(H_a)_{a \in A}$ (où A est une variété) une famille d'opérateurs autoadjoints sur un même espace de Hilbert réel \mathcal{H} , de même domaine D , tous ≥ 0 , à résolvante compacte et dépendant continûment de a au sens que $a \mapsto (Id + H_a)^{-1}$ est continu de A dans $\mathcal{L}(\mathcal{H}, D)$ pour la norme. Si λ est une valeur propre de H_{a_0} de multiplicité N et E_λ^0 l'espace propre, H_a admet pour a voisin de a_0 des valeurs propres proches de λ dont la somme des multiplicités est N , on note E^a la somme des espaces propres associés et on pose $q_\lambda^a = q_{E^a/E_\lambda^0}^a$, où q^a est la forme quadratique sur \mathcal{H} associée à H_a . On note $Q(E)$ l'espace vectoriel des formes quadratiques réelles sur un espace E . On pose la

DÉFINITION. Si $\{\lambda_1 < \lambda_2 < \dots < \lambda_m\}$ est une partie finie du spectre de H , elle est dite **stable** dans la famille $(H_a)_{a \in A}$ si l'application $\Phi: a \mapsto q_{\lambda_1}^a \oplus \dots \oplus q_{\lambda_m}^a$ définie sur un voisinage contractible V de a_0 à valeurs dans $Q(E_{\lambda_1}^0) \oplus \dots \oplus Q(E_{\lambda_m}^0) = \mathcal{E}$ (de dimension l) est une submersion topologique au sens que:

$$\Phi_*: H_{l-1}(V \setminus \Phi^{-1}(e_0); \mathbf{Z}) \rightarrow H_{l-1}(\mathcal{E} \setminus \{e_0\}; \mathbf{Z}) \approx \mathbf{Z}$$

est surjective, où $e_0 = \Phi(a_0)$.

Cette notion est essentielle pour notre problème, car la stabilité a la conséquence suivante: si $\Phi_\varepsilon: V \rightarrow \mathcal{E}$ converge uniformément vers Φ , alors pour $0 < \varepsilon \leq \varepsilon_0$, Φ_ε est stable et donc $e_0 \in \text{Im}(\Phi_\varepsilon)$.

On a en particulier les.

EXEMPLE 1. Soit sur S^2 la famille des laplaciens associés à des métriques conformes à la métrique canonique g_0 , alors pour tout $l \geq 1$ la valeur propre $l(l+1)$ associée à l'espace des harmoniques sphériques de dimension $2l+1$ est stable.

EXEMPLE 2. Sur un tore plat \mathbf{R}^2/Γ toute valeur propre > 0 de multiplicité ≤ 6 est stable dans la famille de tous les laplaciens riemanniens.

Introduisons maintenant la *matrice d'interaction*: soit H un opérateur autoadjoint ≥ 0 sur un espace de Hilbert \mathcal{H}_0 , associé à une forme quadratique fermée q . On note $\mathcal{H}_1 = D(q)$ et \mathcal{H}_{-1} le dual de \mathcal{H}_1 avec les identifications usuelles $\mathcal{H}_1 \xrightarrow{j} \mathcal{H}_0 \xrightarrow{j^*} \mathcal{H}_{-1}$ et $|x|_1^2 = |x|_0^2 + q(x)$, $|\cdot|_{-1}$ la norme duale. On fait les hypothèses suivantes:

(H₁) Les N premières valeurs propres de H vérifient $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N < 1$ et si E désigne la somme des espaces propres associés, $H|_{E^\perp} \geq \text{Id}$.

(H₂) Il existe $E_0 \subset \mathcal{H}_1$ de dimension N tel que: $\forall \varphi \in E_0, |H\varphi|_{-1} \leq \varepsilon |\varphi|_1$.

On a alors le

LEMME (DES PETITES VALEURS PROPRES). *Sous les hypothèses (H_1) et (H_2) , lorsque ε est petit, E et E_0 sont proches et q_E et q_{E_0} sont proches au sens précis suivant: il existe $\varepsilon_0 > 0$ et $C > 0$ ne dépendant que de N tels que si $0 < \varepsilon \leq \varepsilon_0$, on a:*

(a) *si $\mu_1 \leq \dots \leq \mu_N$ sont les valeurs propres de q_{E_0} ,*

$$\mu_i - C\varepsilon^2 \leq \lambda_i \leq \mu_i \quad (\forall i, 1 \leq i \leq N).$$

(b) $\|q_{E/E_0} - q_{E_0}\| \leq C\varepsilon^{1+1/N}$.

Sous les hypothèses H_1 et H_2 , l'espace E_0 muni de la forme quadratique q_{E_0} est donc une excellente approximation de (E, q_E) : on désigne alors la matrice de q_{E_0} dans $(E_0, \langle \rangle)$ sous le nom de *matrice d'interaction*; cette matrice permet grâce au lemme précédent de déterminer le comportement asymptotique précis des petites valeurs propres de H .

3. Spectre de graphes. Soit Γ un graphe (fini pour simplifier l'exposé), non orienté. On désigne par S l'ensemble des sommets et par \mathcal{A} l'ensemble des arêtes. Soit $(V_i)_{i \in S}$ une suite de nombres > 0 et μ la mesure sur S définie par $\mu = \sum_{i \in S} V_i \delta(i)$. Sur l'espace de Hilbert $L^2(S, \mu)$, on introduit une forme quadratique q de la forme $q((x_i)) = \sum_{a \in \mathcal{A}} c_a |d_a x|^2$ où les c_a sont des coefficients > 0 et $d_{\{i,j\}} x = x_i - x_j$. A ces données est associé un *laplacien combinatoire* qui s'écrit

$$(\Delta x)_i = \frac{1}{V_i} \sum_a^i c_a d_a x,$$

où \sum^i est la somme sur les arêtes $a = \{i, j\}$ de sommet i .

EXEMPLE 1 (GRAPHES EN ÉTOILES). $S = \{0, 1, \dots, N\}$ et $\mathcal{A} = \{\{0, i\} | i \geq 1\}$. Désignons par E_N ce graphe, on a la

PROPOSITION. *Soit $0 = \lambda_1 < \lambda_2 < \lambda_3 < \dots < \lambda_{N+1}$ donnés, il existe des $c_i > 0$ ($i \geq 1$) et des $v_i > 0$ ($i \geq 1$) tels que le laplacien associé à $\mu = \delta(0) + \sum_{i=1}^N v_i \delta(i)$ et à la forme quadratique $\sum_{i=1}^N c_i (x_0 - x_i)^2$ admette ces valeurs propres. De plus la partie $\{\lambda_2, \dots, \lambda_{N+1}\}$ de ce spectre est stable relativement aux variations des v_i et c_i .*

EXEMPLE 2 (GRAPHES COMPLETS). On désigne ainsi par C_N le graphe à N sommets tel qu'il existe une arête joignant chaque couple de sommets disjoints.

On a la

PROPOSITION. *Soit $v_i > 0$ ($1 \leq i \leq N$) donnés, pour tout spectre $\lambda_1 = 0 < \lambda_2 \leq \dots \leq \lambda_N$ il existe des $c_{ij} > 0$ ($i \neq j$) tels que ce spectre soit celui associé à mesure $\sum_{i=1}^N v_i \delta(i)$ et à la forme quadratique $\sum_{1 \leq i < j \leq N} c_{ij} (x_i - x_j)^2$. De plus, la partie non nulle de ce spectre est stable relativement aux variations des c_{ij} .*

La preuve utilise une récurrence sur N et la notion de graphe suspendu ST d'un graphe Γ : le graphe suspendu est obtenu en ajoutant un sommet que l'on joint par une arête à tous les sommets de Γ . On peut aussi définir une forme

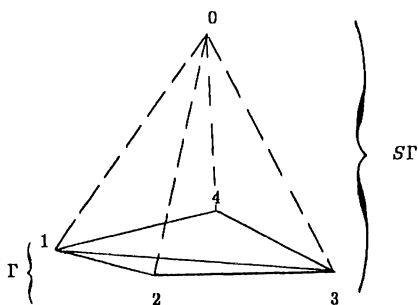


FIGURE 1

quadratique suspendue Sq convenable dont on peut calculer le spectre à partir de celui de q (Figure 1).

4. Effet tunnel géométrique. L'*effet tunnel* s'introduit naturellement dans l'étude de la limite semiclassique ($\hbar \rightarrow 0^+$) de l'opérateur de Schrödinger $-\hbar^2 \Delta + V$ avec des puits de potentiel. La particule quantique a une probabilité très faible, de l'ordre de $\exp(-c/\hbar)$ de traverser les régions classiquement interdites où $V(x) > \text{Energie}$. On peut cependant déterminer avec précision cet effet et appliquer les résultats au splitting des valeurs propres pour les puits symétriques, c'est ce qui est fait dans le travail remarquable [H-S].

Dans un contexte purement géométrique, les barrières de potentiel sont remplacées par des parties étroites de la variété que l'on peut voir comme de *véritables tunnels*! L'espace E_0 sur lequel est calculé la matrice d'interaction est constitué de fonctions de l'espace de Sobolev H^1 , constantes sur les parties larges et harmoniques dans les tunnels.

Décrivons par exemple le cas des *surfaces sans bord*, également déjà entrevu dans [CS]. On considère une surface riemannienne X partagée en domaines $(X_i)_{1 \leq i \leq N}$ dont les bords communs sont des géodésiques périodiques de petites longueurs. On associe à ce découpage le graphe Γ d'adjacence de la famille des X_i (Figure 2).

Les arêtes sont donc en bijection avec les géodésiques périodiques précédentes: $\gamma_a = X_i \cap X_j$ pour $a = \{i, j\}$. Soit l_a la longueur de γ_a . On suppose en outre, ce qui est une hypothèse naturelle lorsque la courbure est constante $\equiv -1$, qu'il existe des voisinages tubulaires Z_a de γ_a de grande largeur $\alpha(a)$ tel que $l_a \cdot \text{ch} \alpha(a) = L > 0$ (L fixée) et la métrique g de X induit sur Z_a la métrique à courbure -1 : $l_a^2 \text{ch}^2 x \cdot d\theta^2 + dx^2$ avec $(x, \theta) \in [-\alpha(a), \alpha(a)] \times \mathbf{R}/\mathbf{Z}$. Soit $\hat{X}_i = X_i \setminus \bigcup_a Z_a$ et E_0 l'espace des fonctions $f \in H^1(X)$ telles que $f|_{\hat{X}_i} = x_i$ et f est harmonique sur les Z_a . Alors, lorsque les $l_a \rightarrow 0^+$, sous l'hypothèse que la première valeur propre > 0 du problème de Neumann des X_i reste minorée par $c > 0$, on peut appliquer le lemme des petites valeurs propres à E_0 . Pour

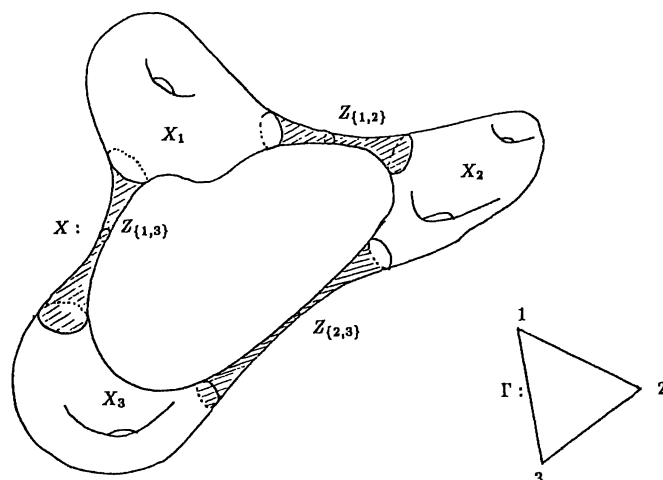


FIGURE 2

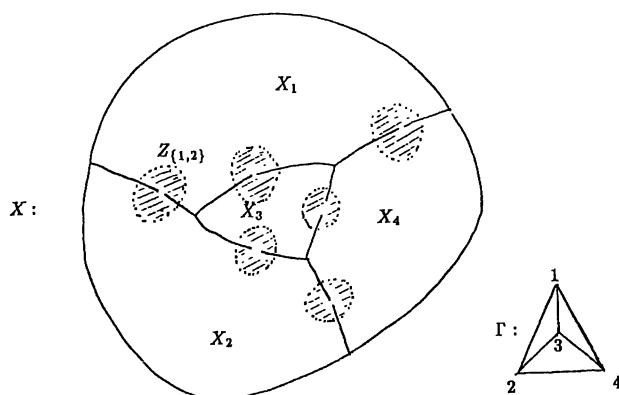


FIGURE 3

$f \in E_0$, on a:

$$\int_X |f|^2 \sim \sum_{i=1}^N V_i x_i^2, \quad \text{avec } V_i = \text{vol}(X_i),$$

et

$$\int_X |df|^2 \sim \frac{1}{\pi} \sum_{a \in \mathcal{A}} l_a |d_a x|^2.$$

On reconnaît bien sûr un laplacien combinatoire associé au graphe Γ . Mentionnons en passant que la méthode s'applique aussi à l'évaluation des petites valeurs propres d'une surface d'aire infinie à courbure -1 ([GL] et aussi [P-S]).

On peut faire une construction voisine pour un ouvert euclidien X de \mathbf{R}^2 , $X = \bigcup_{i=1}^N X_i$, où $X_i \cap X_j$ est lorsque $\{i, j\} \in \mathcal{A}$ un petit intervalle rectiligne de longueur $2\varepsilon_a$ (Figure 3).

Les Z_a sont alors des disques centrés au milieu de ces intervalles et de rayon $\rho > 0$ fixé lorsque les $\varepsilon_a \rightarrow 0^+$. La construction de E_0 est la même que pour les surfaces sans bord et l'évaluation donne:

$$\int_X |df|^2 \sim \frac{\pi}{2} \sum_{a \in \mathcal{A}} \frac{1}{|\text{Log}(\varepsilon_a)|} |d_a x|^2.$$

5. Perturbations singulières. Pour prouver les autres énoncés (variétés de dimension ≥ 3 , propriété (C_N) pour les surfaces), on procède par perturbations singulières. Il faut bien entendu dans chaque cas prouver des résultats plus forts que la seule convergence des valeurs propres afin de pouvoir utiliser les propriétés de stabilité: on doit avoir convergence uniforme de certaines formes quadratiques.

Cas des variétés de dimension ≥ 3 . Pour munir une telle variété X d'une métrique telle que les N premières valeurs propres du laplacien soient la suite $s_N = \{\lambda_1 = 0 < \lambda_2 \leq \dots \leq \lambda_N\}$ donnée, on commence par choisir une surface S de genre assez grand pour que $N(S) \geq N$. On choisit alors une métrique g_0 sur S ayant s_N comme suite des N premières valeurs propres du laplacien et de façon stable relativement à une famille de métriques $g \in K$ voisines de g_0 . On plonge alors S dans X et on choisit pour chaque $g \in K$ (continûment par rapport à g) une métrique G_g sur X telle que S admette un voisinage isométrique à $[-2a, +2a] \times S$ muni de la métrique $dt^2 + g$. Lorsque a est assez petit le domaine $D_a = [-a, +a] \times S$ admet pour la métrique G_{g_0} et le problème de Neumann, la suite s_N comme début du spectre et de façon stable dans K .

Il reste à construire une suite de métriques sur X dont le spectre converge vers celui de D_a . Une possibilité est la suivante (et $d \geq 3$ intervient aussi ici): soit $G_g^{n,\varepsilon}$ la métrique conforme à G_g donnée par $G_g^{n,\varepsilon} = \varphi_{n,\varepsilon} G_g$ où $\varphi_{n,\varepsilon} \in C^\infty(X; \mathbb{R})$,

$$\varphi_{n,\varepsilon} \equiv \varepsilon \text{ sur } X \setminus D_{a+(1/2n)}, \quad \varphi_{n,\varepsilon} \equiv 1 \text{ sur } D_{a+(1/n)}$$

et $\varepsilon \leq \varphi_{n,\varepsilon} \leq 1$ partout. On peut construire une suite $G_g^{n(\varepsilon),\varepsilon}$ dont les valeurs propres convergent vers celle de (D_a, G_g) .

Case des surfaces. Pour la sphère S^2 la propriété (C_N) est conséquence immédiate de la stabilité des harmoniques sphériques: on choisit l tel que $2l+1 \geq N$ et on considère sur l'espace propre \mathcal{H}_l une forme quadratique q_ε de valeurs propres

$$l(l+1) + \varepsilon \mu_1 \leq \dots \leq l(l+1) + \varepsilon \mu_M < \dots, \quad \text{pour } \varepsilon \text{ assez petit,}$$

il existe g proche de la métrique usuelle de S^2 dont les valeurs propres proches de $l(l+1)$ sont celles de q_ε . La métrique $\sqrt{\varepsilon}g$ permet de vérifier la propriété (C_N) .

Pour obtenir la même propriété sur d'autres surfaces, on utilise l'adjonction de petites anses (orientables ou non) et les résultats de [C-F] et [A-E]: soit S une surface riemannienne et $(x_i)_{1 \leq i \leq 2n}$ une suite de points de S deux à deux distincts. On se donne des $l_i > 0$ ($1 \leq i \leq n$) et on construit la surface S_ε en recollant à $S \setminus \bigcup_{i=1}^{2n} B(x_i, \varepsilon)$ des anses euclidiennes de longueur l_i et de rayon ε .

Lorsque $\varepsilon \rightarrow 0$, le spectre du laplacien sur S_ε (métrique C^1 par morceaux) converge vers la réunion du spectre de S et des spectres de Dirichlet des intervalles $[0, l_i]$.

RÉFÉRENCES

- [AD] V. Arnold, *Modes and quasi-modes*, J. Funct. Anal. **6** (1972), 94–101.
- [AE] C. Anné, *Spectre du laplacien et limites de variétés avec perte de dimension I*, (à paraître, 1985).
- [BN] G. Besson, *Sur la multiplicité de la première valeur propre des surfaces riemanniennes*, Ann. Inst. Fourier (Grenoble) **30** (1980), 109–128.
- [BR] M. Burger, *Estimations des petites valeurs propres du laplacien d'un revêtement de variétés riemanniennes compactes*, C. R. Acad. Sci. Paris Sér. I Math. **302** (1986), 191–194.
- [BS1] R. Brooks, *The first eigenvalue in a tower of coverings*, Bull. Amer. Math. Soc. (N.S.) **13** (1985), 137–140.
- [BS2] ———, *The spectral geometry of a tower of coverings*, Preprint, 1985.
- [C-F] I. Chavel et A. Feldmann, *Isoperimetric constants of manifolds with small handles*, Math. Z. **184** (1983), 435–448.
- [CS] B. Colbois, *Petites valeurs propres du laplacien sur une surface de Riemann compacte et graphes*, C. R. Acad. Sci. Paris Sér. I Math. **301** (1985), 927–930.
- [CV] Y. Colin de Verdière, *Sur la multiplicité de la première valeur propre non nulle du laplacien*, Comment. Math. Helv. **61** (1986), 254–270.
- [CG] S. Cheng, *Eigenfunctions and nodal sets*, Comment. Math. Helv. **51** (1979), 43–45.
- [DK] J. Dodziuk, *Difference equations, isoperimetric inequalities and transience of certain random walks*, Trans. Amer. Math. Soc. **284** (1984), 787–794.
- [GL] P. Gall, *Sur la première valeur propre des groupes de Hecke de covolume presque fini*, C. R. Acad. Sci. Paris Sér. I Math. **302** (1986), 299–302.
- [H-S] B. Helffer et J. Sjöstrand, *Puits multiples en semi-classique*. I, Comm. Partial Differential Equations **9** (1984), 337–408; II, Ann. Inst. H. Poincaré Phys. Théor. **42** (1985), 127–212; III, Math. Nachr. **127** (1985), 263–313; IV, Comm. Partial Differential Equations **10** (1985), 245–340; V, VI (à paraître).
- [KC] M. Kac, *Can one hear the shape of a drum?*, Amer. Math. Monthly **73** (1966), 1–23.
- [P-P-W] L. Payne, G. Polya, et H. Weinberger, *On the ratio of consecutive eigenvalues*, J. Math. Phys. **35** (1956), 289–298.
- [PR] M. Protter, *Can one hear the shape of a drum?*, Revisited, Preprint, Berkeley, 1985.
- [P-S] T. Pignatoro et D. Sullivan, *Ground state and lowest eigenvalue of the Laplacian for non-compact hyperbolic surfaces*, Comm. Math. Phys. **104** (1986), 529–535.
- [RA] R. Rammal, *Spectrum of harmonic excitations on fractals*, J. Physique **45** (1984), 191–206.
- [RL] G. Ringel, *Map color theorem*, Springer, 1974.

INSTITUT FOURIER, 38402 SAINT MARTIN D'HÈRES CEDEX, FRANCE

Combinatorial Methods in Symplectic Geometry

YA. M. ELIASHBERG

Graphs arise when studying many problems of symplectic and contact topology. Sometimes the structure of the graphs can be investigated and this delivers deep results towards “rigidity” of contact and symplectic structures. In this report I shall discuss two such combinatorial reductions in the simplest situation.

1. The structure of one-dimensional wave fronts.

1.1. Let us consider \mathbf{R}^3 as a standard contact space $\mathcal{J}^1(\mathbf{R}^1)$ with coordinates (x, y, z) and contact form $dz - ydx$. Denote by π the projection $\mathbf{R}^3 = \mathcal{J}^1(\mathbf{R}^1) \rightarrow \mathcal{J}^0(\mathbf{R}^1) = \mathbf{R}^2$ which is determined by the formula $\pi(x, y, z) = (x, z)$ and by p the projection $(x, z) \mapsto x$ of $\mathbf{R}^2 = \mathcal{J}^0(\mathbf{R}^1)$ onto \mathbf{R}^1 . A *wave front* in \mathbf{R}^2 is the projection by π of an oriented legendrian curve in \mathbf{R}^3 . It appears as a (possibly self-intersecting) plane curve with cusp points (see Figure 1). A *branch* of the front is a part between two neighboring cusp points. Each branch can be considered as the graph $z = \varphi(x)$ of a partly defined function φ on \mathbf{R}^1 . A *generic* front F has branches that intersect transversally, has no triple intersection points, and has no cusp points which are double points of F . Henceforth we consider only generic fronts and only two types of them: closed fronts and compact fronts with boundary. In the last case we require boundary points not to be double points of the front.

The oriented *graph* G_F (abstract or immersed in \mathbf{R}^2) of the front F is defined as follows. For its vertices we take cusp points of F and boundary points of F if F is not closed. Edges of G_F correspond to branches of F (see Figure 2). The orientation of an edge is induced from the orientation of the original legendrian curve by π . The edge is called *positively* (*negatively*) oriented if its orientation is preserved (reversed) by the projection p . Let $\mathcal{P} = \{p_1, \dots, p_s\}$ be a set of (not necessarily all) self-intersection points of F . Denote by $G_F^{\mathcal{P}}$ the graph received from G_F by adding new vertices p_1, \dots, p_s and subdividing its oriented edges. An oriented path in $G_F^{\mathcal{P}}$ is called a *semicycle* if it consists of edges of the same sign. A closed path in $G_F^{\mathcal{P}}$ is called a *cycle* if the branches of F corresponding to its edges do not intersect and if it contains between its vertices exactly two cusp points which divide it into two semicycles (see Figure 3). A front F is called

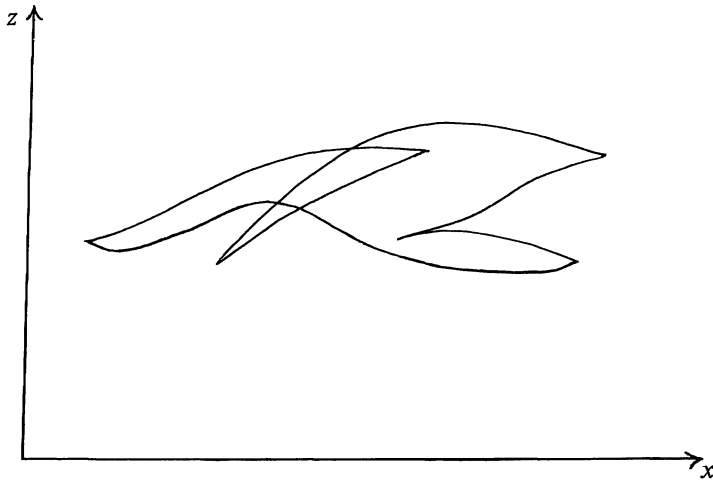


FIGURE 1

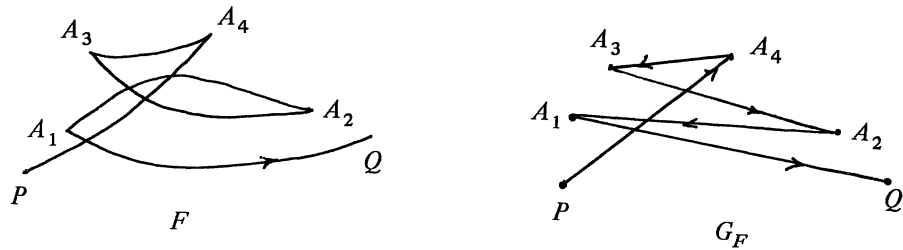


FIGURE 2. A front F and its graph G_F .

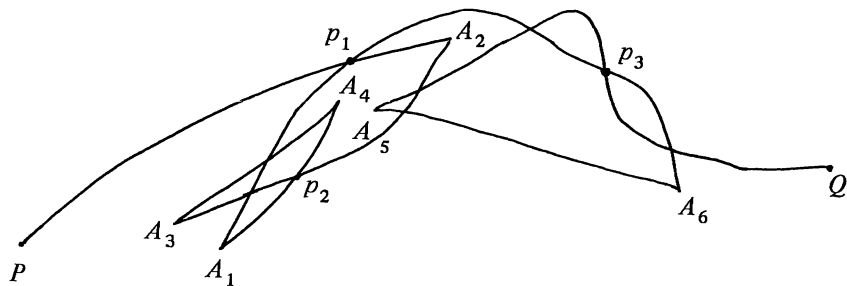


FIGURE 3. A desintegratable front F . The graph $G_F^{\mathcal{P}}$ for $\mathcal{P} = \{p_1, p_2, p_3\}$ can be presented as the union of the semicycle Pp_1p_3Q and 3 cycles: $p_1A_2A_1$, $p_2A_3A_4$, and $p_3A_6A_5$.

desintegratable if for a set $\mathcal{P} = \{p_1, \dots, p_s\}$ the graph $G_F^{\mathcal{P}}$ can be presented as the union of cycles and possibly one semicycle ending at boundary points of F (see Figure 3). A homotopy F_t , $t \in [0, 1]$, between two fronts F_0 and F_1 is called *transversal* if it is covered by an isotopy of legendrian curves and for all $t \in [0, 1]$ boundary points of F_t are left single by the homotopy. Note that for any $t \in [0, 1]$ branches of F_t have a transversal intersection.

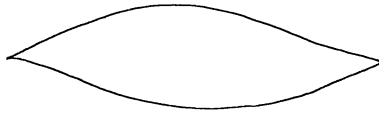


FIGURE 4. The standard front.

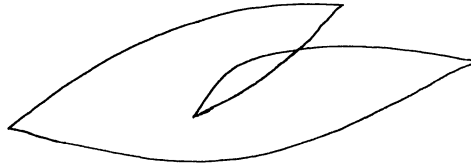


FIGURE 5. A nonstandard front.

The fronts presented in Figure 4 and legendrian curves covering them we call standard. The main front structure theorem in the situation under consideration can be formulated as follows.

THEOREM A. *If a wave front is transversally homotopic to the standard front then it is desintegratable.*

Theorem A, its higher-dimensional analogs and applications are discussed in my paper [1]. The complete proof of the theorem is not published.

1.2. Corollaries A1, A2, and A3 in this section illustrate the character of applications of Theorem A.

Consider the legendrian curve $L \subset \mathbf{R}^3$ with the front presented in Figure 5. This curve is evidently unknotted and legendrian regularly homotopic to the standard closed curve but its front is not desintegratable. Thus we have

COROLLARY A1. *There exists an unknotted legendrian curve in \mathbf{R}^3 which is regularly legendrian homotopic to the standard curve but cannot be connected with the standard curve by a legendrian isotopy.*

Let Leg be the space of legendrian embeddings $\varphi: [0, 1] \rightarrow \mathcal{I}^1([0, 1]) \subset \mathcal{I}^1(\mathbf{R}^1) = \mathbf{R}^3$ with $p(\pi(\varphi(i))) = i$, $i = 0, 1$. By Leg_0 we denote the component of Leg containing the standard legendrian curve.

COROLLARY A2. *Let $\varphi \in \text{Leg}_0$ and $\varphi([0, 1]) \subset \{y > 0\} \subset \mathbf{R}^3$. Let the boundary points $P = \pi(\varphi(0))$ and $Q = \pi(\varphi(1))$ of the front of φ have coordinates $(0, z_0)$ and $(1, z_1)$. Then $z_1 > z_0$.*

PROOF. By Theorem A the front of φ is desintegratable. Thus the graph G_F^P for some set $P = \{p_1, \dots, p_s\}$ contains a semicycle connecting P and Q (see Figure 6). It corresponds to the graph of a piecewise smooth function $h: [0, 1] \rightarrow \mathbf{R}$ with $h(0) = z_0$, $h(1) = z_1$. The condition $\varphi([0, 1]) \subset \{y > 0\}$ gives the inequality $h'(x) > 0$ for $x \in]0, 1[$ and hence $z_1 = h(1) > h(0) = z_0$.

Let now K be the square $\{0 \leq x, z \leq 1\}$ in the plane $\{y = 0\} \subset \mathbf{R}^3$ and let $S \subset \mathcal{I}^1([0, 1]) \subset \mathbf{R}^3$ be a surface with $\partial S = \partial K$ tangent to K at ∂S (see Figure 7). We call S simple if it is transversal to the contact distribution τ .

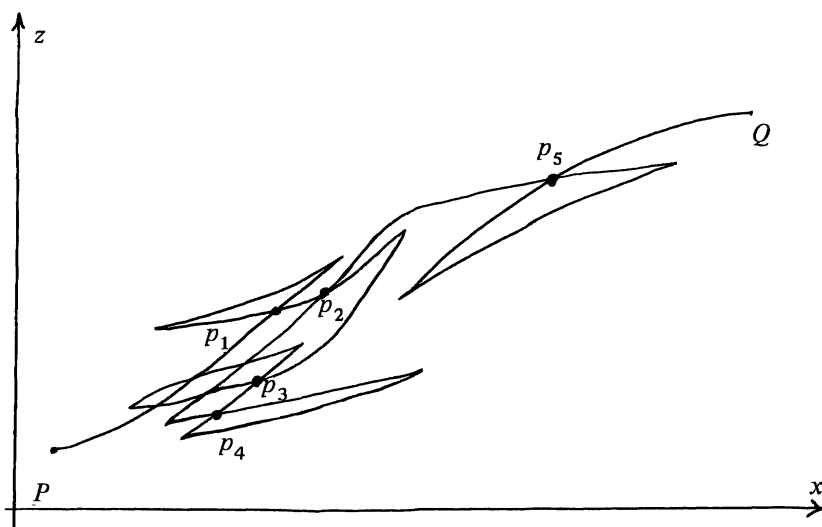


FIGURE 6. The semicycle $Pp_1p_2p_3p_4p_5Q$ corresponds to the graph of a piecewise smooth function.

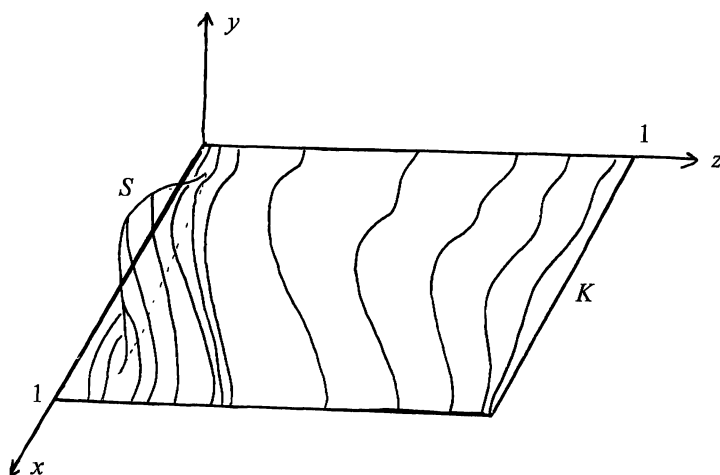


FIGURE 7

Integral curves of $\tau \cap S$ (and $\tau \cap K$) are legendrian curves with their ends at $I_i = \{x = i, y = 0, 0 \leq z \leq 1\}$, $i = 0, 1$, and define the diffeomorphism of holonomy

$$f_S: [0, 1] \rightarrow [0, 1].$$

Applying Corollary A2, we get

COROLLARY A3. *If S does not intersect K at interior points then f_S has no interior fixed points.*

Corollary A3 is a weak version of Bennequin's theorem (see [3]) but it is sufficient to recognize nonstandard contact structures on \mathbf{R}^3 .

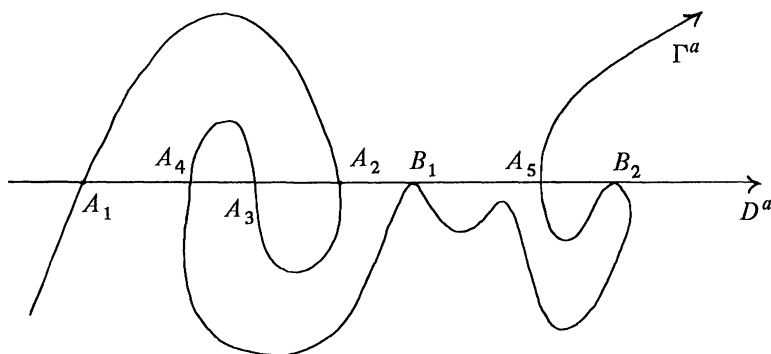


FIGURE 8. A_1, A_3, A_5 —crossings; A_4, A_2 —passages; B_1 —a positive critical point; B_2 —a negative critical point.

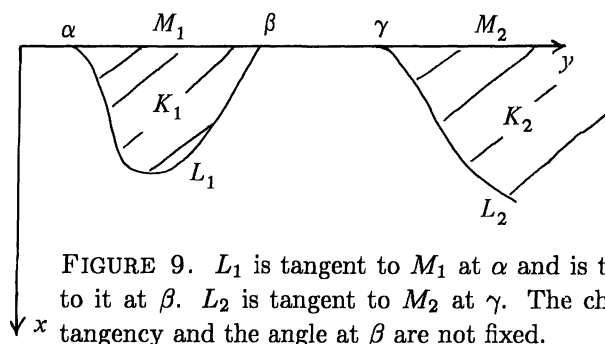


FIGURE 9. L_1 is tangent to M_1 at α and is transversal to it at β . L_2 is tangent to M_2 at γ . The character of tangency and the angle at β are not fixed.

2. The estimates on the number of fixed points of area-preserving transformations.

2.1. Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be an orientation-preserving diffeomorphism and $S = (x, y, z)$ be a coordinate system in $\mathbf{R}^3 = \mathbf{R}^2 \times \mathbf{R}^1$. Let f_1, f_2 be the coordinate function of f , Γ the graph $\{z = f_1(x, y)\} \subset \mathbf{R}^3$ of f_1 and D the diagonal $\{z = x\} \subset \mathbf{R}^3$. D divides \mathbf{R}^3 into two semispaces $\mathbf{R}_+^3 = \{z \geq x\}$ and $\mathbf{R}_-^3 = \{z \leq x\}$. Let π and p be the projections of $\mathbf{R}^3 = \mathbf{R}^2 \times \mathbf{R}^1$ on the first and second factors and T be the intersection $\Gamma \cap D$. For $a \in \mathbf{R}$ let $\Pi^a = \{z = a\}$, $D^a = D \cap \Pi^a$, $\Gamma^a = \Gamma \cap \Pi^a$.

The orientation of D, Γ, Π^a is induced from \mathbf{R}^2 by π . We orient D^a and Γ^a as boundaries of oriented surfaces $D \cap \{z \geq a\}$ and $\Gamma \cap \{z \geq a\}$.

Suppose now that Γ is transversal to D and the restriction $p|_T: T \rightarrow \mathbf{R}$ is the Morse function with different critical values. This is a generic property of the diffeomorphism f .

Let C be the set of critical points of $p|_T$. A critical point $v \in C$ corresponding to a critical value $a \in \mathbf{R}$ is called *positive* (resp. *negative*) if the manifolds Γ^a and D^a have the same (resp. opposite) orientation at v .

A point $v \in (T \setminus C) \cap \Pi^a$ is called a *crossing* (resp. a *passage*) if points of $\mathbf{R}_-^3 \cap \Gamma^a$ near v precede (resp. follow) v in the sense of the orientation of Γ^a (see Figure 8).

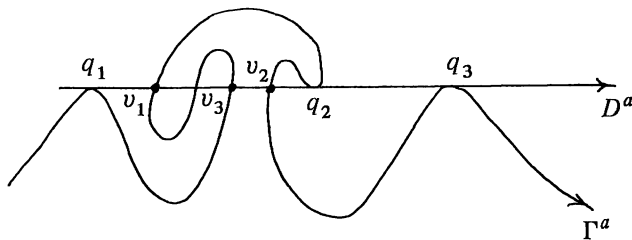


FIGURE 10. v_1, v_3 are neighbors of q_1 ; v_2 is a neighbor of q_2 ; q_3 has $+\infty_y$ as its neighbor.

Let $K_i, L_i, M_i \subset \mathbf{R}^2$, $i = 1, 2$, be the sets presented in Figure 9. We have $\partial K_i = L_i \cup M_i$, $i = 1, 2$, $M_1 \cap L_1 = \{\alpha, \beta\}$, and $M_2 \cap L_2 = \{\gamma\}$. M_i and L_i are oriented as parts of boundaries of K_i , $i = 1, 2$.

Let $q \in C \cap \Pi^a$. A crossing $v \in T^a$ is called a *neighbor* of q if there exists an immersion $\Phi: K_1 \rightarrow \Pi^a$ with $\Phi(\alpha) = q$, $\Phi(\beta) = v$, $\Phi(M_1) \subset D^a$, $\Phi(L_1) \subset \Gamma^a$.

Note that Φ preserves orientation if q is positive and reverses it otherwise.

We say that a positive critical point $q \in C \cap \Pi^a$ has $+\infty_y$ (resp. $-\infty_y$) as its neighbor if there exists a proper orientation-preserving immersion $\Phi: K_2 \rightarrow \Pi^a$ with $\Phi(\gamma) = q$, $\Phi(M_2) \subset D^a$, $\Phi(L_2) \subset \Gamma^a$ and such that the restriction $\Phi|_{M_2}: M_2 \rightarrow \Gamma^a$ preserves (resp. reverses) orientation and the restriction $\Phi|_{L_2}: L_2 \rightarrow D^a$ reverses (resp. preserves) it (see Figure 10).

Let $\sigma: T \rightarrow \mathbf{R}$ be the function defined by the formula $\sigma(x, y, x) = f_2(x, y) - y$. Note that zeros of σ correspond to fixed points of the diffeomorphism f .

Suppose now that f has only isolated fixed points.

Denote by V , $V \subset T$, the closure of the set of crossings. Then V is a one-dimensional manifold with the boundary C . An infinite (abstract or immersed in D) graph G is defined as follows. The set of vertices of G consists of points of C , their neighbors, zeros of the function σ and possibly four special points: $\pm\infty_x$, $\pm\infty_y$. There are two types of edges of G : *vertical* and *horizontal*. The set of vertices divides V into several arcs. Vertical edges of G correspond to these arcs. A noncompact arc without boundary points defines the edge connecting vertices $-\infty_y$ and $+\infty_x$. A noncompact arc l with one boundary point q defines the edge $+\infty_x q$ if $p(v) \geq p(q)$ for all $v \in l$ and the edge $-\infty_x q$ if the opposite inequality holds. Horizontal edges connect vertices of C with their neighbors (including $\pm\infty_y$). Edges ending at points $\pm\infty_x, \pm\infty_y$ are called rays. The orientation of edges of G is defined as follows. If the horizontal edge connects a point $q \in C$ with its neighbor v we choose the direction from q to v if q is positive and from v to q if q is negative. Note that rays $\pm\infty_y q$ are always oriented from q to $\pm\infty_y$. A vertical edge qv with $p(q) < p(v)$ we orient from q to v if the function σ is positive on the corresponding arc of V and from v to q otherwise.

An oriented path in G is called a *semicycle* if it is infinite or ends by a ray (see Figure 11).

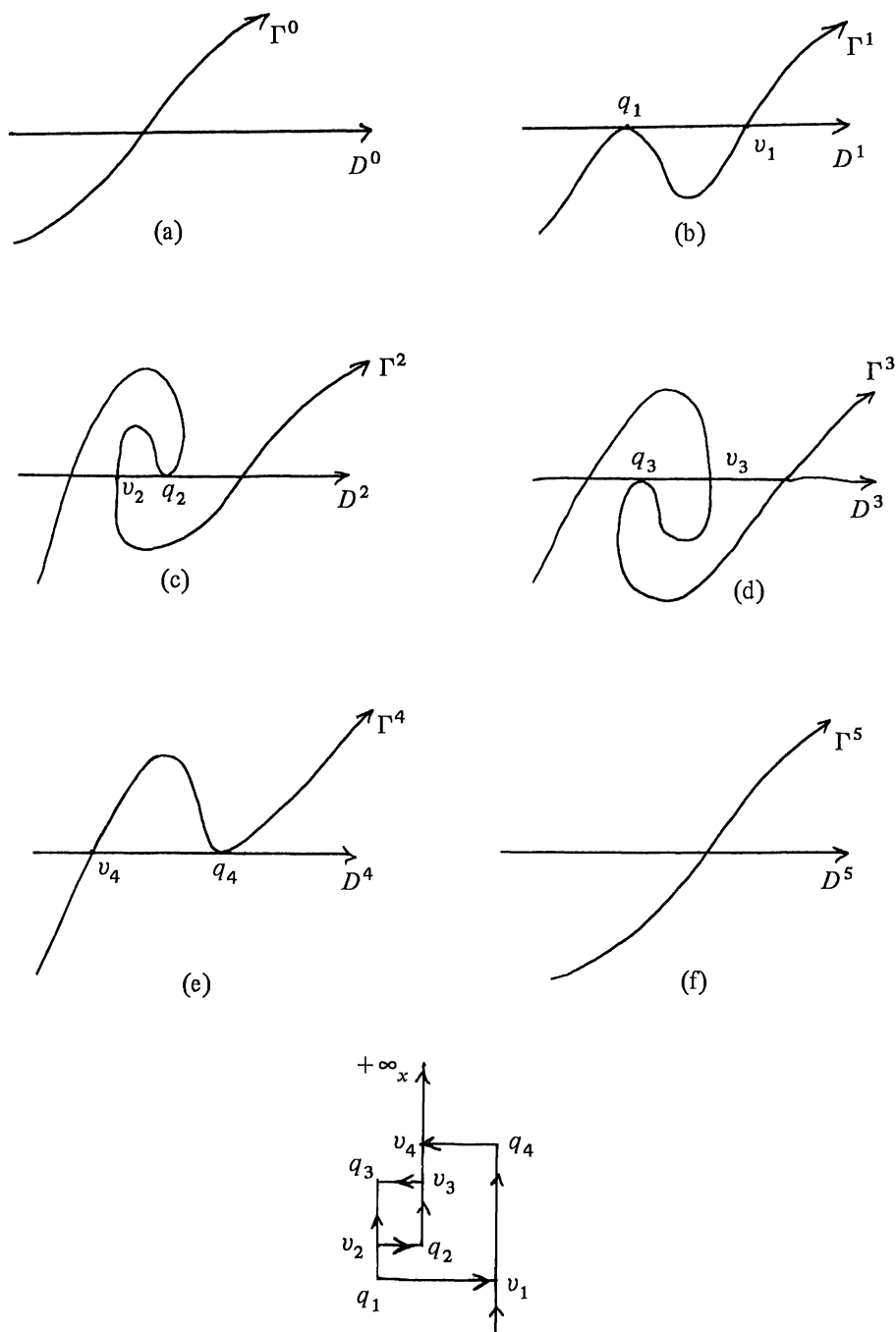


FIGURE 11. The graph G for the diffeomorphism f . $v_1 q_4 v_4 + \infty_x$ —a semicycle. In (a)–(f) a diffeomorphism f is presented by manifolds Γ^a for $a = 0, 1, \dots, 5$.

The definition of G depends on the choice of the coordinate system $S = (x, y, z)$. The coordinate change $(x, y, z) \mapsto (x, y, -z)$ leads to another graph which we denote by G^* . Note that vertical edges of G^* correspond to arcs of the manifolds of passages of the original construction.

The orientation-preserving diffeomorphism f is called *S-limited* if

$$\lim_{y \rightarrow \pm\infty} f_2(x, y) = \pm\infty$$

for all $x \in \mathbf{R}$.

The following theorem was proved in my paper [2].

THEOREM B. *Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be an S-limited diffeomorphism without fixed points of index +1. If $T \neq \emptyset$ then at last one of the graphs G and G^* has a semicycle.*

2.2. Going back to the definition of the semicycle it is easy to construct under the conditions of Theorem B a path in \mathbf{R}^2 bounding with its image by f an immersed band in \mathbf{R}^2 . Strictly speaking we have

COROLLARY B1. *Under the conditions of Theorem B there exists a proper immersion $h: [0, +\infty[\times [0, 1] \rightarrow \mathbf{R}^2$ such that $f(h(t, 0)) = h(t, 1)$ for $t \in [0, +\infty[$, and $f_1(h(0, u)) = f_1(h(0, 0))$ for $u \in [0, 1]$.*

Let now M be a closed oriented surface of the genus $g > 0$ and $\varphi: M \rightarrow M$ be a diffeomorphism without fixed points of index +1 which is homotopic to the identity mapping $\text{id}: M \rightarrow M$. Choosing carefully a coordinate system on the covering plane \mathbf{R}^2 , it is easy to satisfy the conditions of Theorem B for the covering transformation $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$. Applying Theorem B and using the compactness of M we get

COROLLARY B2. *Let $\Phi: M \times [0, 1] \rightarrow M$ be a homotopy between id and φ . There exists an immersion $h: S^1 \times [0, 1] \rightarrow M$ with $\varphi(h(t, 0)) = h(t, 1)$ for $t \in [0, 1]$ which is homotopic $\text{rel } \partial(S^1 \times [0, 1])$ to the mapping $(t, u) \mapsto \Phi(h(t, 0), u)$.*

Note that for $g > 1$ the last part of the assertion is automatic.

Suppose now that the surface M is supplied by a volume form ω and that the diffeomorphism φ preserves ω . Denote by $\Delta_\varphi^\omega, \Delta_\varphi^\omega \in H^1(M; K)$, the "Calabi class" (see [2, 5]) of φ ($K = \mathbf{R}/\mathbf{Z}$ if M is the torus and $K = \mathbf{R}$ in other cases). Recall that the value of Δ_φ^ω on a cycle Z is equal to the area covered up by Z moving by the action of a homotopy between id and φ . The next corollary immediately follows from B2.

COROLLARY B3. *Let $\varphi: M \rightarrow M$ be a diffeomorphism which preserve the form ω and is homotopic to the identity mapping $\text{id}: M \rightarrow M$. If $\Delta_\varphi^\omega = 0$ then φ has at least one fixed point of index +1.*

In fact the assertion of B3 applied to an auxiliary surface (see [2]) gives for the original diffeomorphism at least two fixed points of index +1 which implies well-known estimates of the minimal number of fixed points of φ . These estimates

were originally proved in [2] in 1979 and were recently obtained in [4] by the Conley-Zehnder variational method.

REFERENCES

1. Ya. M. Eliashberg, *The wave front structure theorem. I*, Funktsional. Anal. i Prilozhen. (1986), in press.
2. ———, *Estimates on the number of fixed points of area preserving transformation*, Syktyvkar, 1979.
3. D. Bennequin, *Entrelacements et équations de Plaff*, Astérisque **107–108** (1982), 87–162.
4. J.-C. Sikorav, *Points fixe d'une application homologue à l'identité*, J. Differential Geom. **22** (1985), 49–79.
5. E. Calabi, *On the group of automorphisms of a symplectic manifold*, Problems in Analysis (Symposium in honor of S. Bochner), Princeton Univ. Press, Princeton, N.J., 1970, pp. 1–26.

INSTITUTE OF ACCOUNTING, LENINGRAD, USSR

Singularities in Some Geometric Variational Problems

ROBERT M. HARDT

1. Introduction. Singularities occur in solutions for a wide variety of nonlinear variational problems that arise in geometry. Here we will, for two particular classes of problems, analyze several questions which are associated in general with the occurrence of singularities. The first class of problems involves mappings between Riemannian manifolds that are critical points for a suitable elliptic energy functional. The second class involves rectifiable currents that are critical points for a parametric area functional.

For simplicity, we will assume that M is a smooth bounded open subset of \mathbb{R}^m and that N is a compact smooth n -dimensional submanifold of \mathbb{R}^{n+k} . We will compare the *least energy mapping problem*:

Given a smooth $\varphi: \partial M \rightarrow N$, find a least energy $u: M \rightarrow N$ with $u|_{\partial M} = \varphi$,
with the *least area problem*:

Given a compact smooth oriented $(n-1)$ -dimensional submanifold Γ of \mathbb{R}^{n+k} , find an n -dimensional rectifiable current T of least area with $\partial T = \Gamma$.

In the first problem, we will mainly discuss, for $1 < p < \infty$, the p -energy $\int_M |\nabla u|^p dx$, the most important case being $p = 2$ where critical points are *harmonic mappings*. In the second problem, a rectifiable current T is simply a countable sum of disjoint pieces N_i of oriented n -dimensional C^1 submanifolds with integer multiplicities m_i . The area of T is $\sum |m_i| \mathcal{H}^n(N_i)$ (where \mathcal{H}^n denotes n -dimensional Hausdorff measure), and the equation $\partial T = \Gamma$ means that $\int_\Gamma \psi = \sum m_i \int_{N_i} d\psi$ for all smooth $n-1$ forms ψ . (Rectifiable currents were introduced in [FF] and discussed extensively in [F₁]; other useful references include [F₃] and [S₂].)

2. Singularities by topological obstruction. Topological obstructions may be relevant for the existence or regularity of minimizers. Perhaps the simplest example concerns the case of maps from the unit ball \mathbb{B}^m in \mathbb{R}^m to the

Research partially supported by the National Science Foundation.

sphere $\mathbb{S}^{m-1} = \partial\mathbb{B}^m$. Here an elementary topological condition is

$$\begin{aligned} &\text{there exists a continuous map } u: \overline{\mathbb{B}}^m \rightarrow \mathbb{S}^{m-1} \\ &\text{if and only if } u|_{\partial\mathbb{B}^m} \text{ has degree 0.} \end{aligned} \quad (*)$$

For $p \geq m$, $(*)$ implies that there exists no function $u: \overline{\mathbb{B}}^m \rightarrow \mathbb{S}^{m-1}$ of finite p -energy whose trace $u|_{\partial\mathbb{B}^m}$ has nonzero degree.

In fact, for $p > m$, a finite energy u would be essentially continuous by Sobolev embedding while for $p = m$, one could suitably approximate u by a continuous map to contradict $(*)$. Thus for $p \geq m$, the least energy boundary-value problem may be meaningless.

On the other hand, if $1 < p < m$, then there is such a finite p -energy extension given, for example, by the homogeneous-degree-0 extension, $w(x) = \varphi(x/|x|)$. In this case, the boundary-value problem is meaningful, but $(*)$ implies that a solution will necessarily be discontinuous on $\overline{\mathbb{B}}^m$ if the given φ has nonzero degree (e.g., $\varphi = \text{identity}$).

In general, if there exists some finite p -energy extension of φ , then one easily obtains the existence of a solution of the least p -energy boundary value problem in the class

$$\mathcal{U}_\varphi = L^{1,p}(M, \mathbb{R}^{n+k}) \cap \{u: u|_{\partial M} = \varphi \text{ and } u(x) \in N \text{ for a.e. } x\}$$

where $L^{1,p}(M, \mathbb{R}^k)$ is the space of functions of finite p -energy [KJF].

A topological obstruction may also lead to singularity in the least area problem. An area minimizing rectifiable current T must have a singularity if $\Gamma = \partial T$ determines a nonzero cobordism class, for example, if Γ is a $\mathbb{C}P^2$ embedded in \mathbb{S}^4 .

3. Singularities without topological obstructions. In [HL₁] is a specific example of a smooth map $\varphi: \mathbb{S}^{m-1} \rightarrow \mathbb{S}^{m-1}$ of degree 0 for which there is a definite gap between the two numbers

$$\inf \left\{ \int_{\mathbb{B}^m} |\nabla u|^2 dx : u \in \mathcal{U}_\varphi \right\} < \inf \left\{ \int_{\mathbb{B}^m} |\nabla v|^2 dx : v \in \mathcal{U}_\varphi \cap C^0(\overline{\mathbb{B}}^m, \mathbb{S}^{m-1}) \right\}.$$

In particular, the energy minimizer must have a singularity even though there is no topological restriction in the sense that there does exist some continuous finite-energy extension of φ . The gap may be made arbitrarily large by suitable choice of φ . To explain the idea of this construction we will describe an analogous problem where $m = 3$ and \mathbb{B}^3 is replaced by a spatial region M shaped like a barbell with a thin handle of length L . The boundary data φ is given by a unit vectorfield on ∂M , as shown in Figure 1.

As a map from ∂M to \mathbb{S}^2 , φ has degree 0. By considering a comparison function that is approximately constant on the handle and homogeneous of degree 0 on each end, we readily check that

$$\inf \left\{ \int_M |\nabla u|^p dx : u \in \mathcal{U}_\varphi \right\} < 2(8\pi) + 1.$$

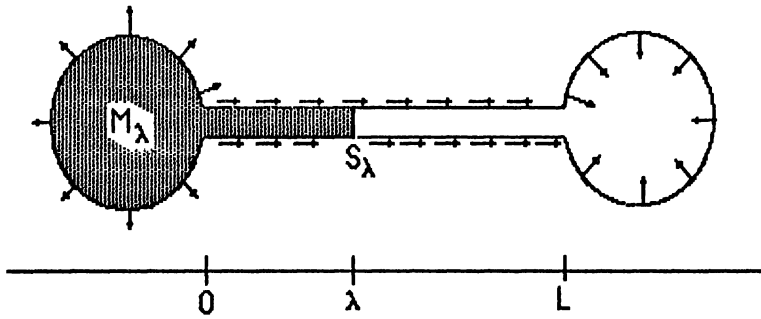


FIGURE 1

On the other hand, suppose that $v \in \mathcal{U}_\varphi \cap C^0(\overline{\mathbb{B}^3}, \mathbb{S}^2)$. Then, for all $0 < \lambda < L$, $v|_{\partial M_\lambda}$ has degree 0 where M_λ is the subregion of M shown above. Then the restriction of v to the vertical slice S_λ must cover more than half of \mathbb{S}^2 because $v|_{\partial M_\lambda} \sim S_\lambda = \varphi|_{\partial M_\lambda} \sim S_\lambda$ almost covers all of \mathbb{S}^2 . Since the energy of a map of a surface is at least twice the area of the map, this gives a lower bound for the energy of the slice:

$$\int_{S_\lambda} |\nabla v|^2 d\mathcal{H}^2 \geq \int_{S_\lambda} |\nabla_{\tan} v|^2 d\mathcal{H}^2 \geq 2 \text{Area}(v|_{S_\lambda}) \geq 4\pi.$$

By Fubini's theorem, $\int_M |\nabla v|^p dx \geq 4\pi L \rightarrow \infty$ as $L \rightarrow \infty$.

For the least area problem, singularities may also occur without a topological obstruction. For example, the intersection of a holomorphic variety in \mathbb{C}^2 (e.g., $\{(y, z): z^2 = y^3\}$) with the unit ball defines, by [F₁, 5.4.19], an area-minimizing rectifiable current T even though ∂T may be a smooth curve that bounds some embedded surface. Similarly, in codimension one, the cone $0 \times (\mathbb{S}^3 \times \mathbb{S}^3)$ is area-minimizing [BDG] even though $\mathbb{S}^3 \times \mathbb{S}^3$ bounds some smooth embedded 7-manifold.

4. Partial regularity theory. Having seen that singularities in solutions are often unavoidable, we now discuss estimations on their size. There are several parallels in the proofs. Both often involve a regularity result for minimizers having sufficiently small energy or area excess and an approximation by harmonic functions or by graphs of harmonic functions. Least area regularity theory has an additional complexity in that the minimizing current T is not given as the image or graph of a function, and another initial graphical approximation result is needed. Nevertheless, most work on regularity for the least area problem preceded that for the least energy problem. First we will give a brief summary of what is known.

For the least area problem, one may view the fundamental rectifiable compactness theorem of [FF] (see also [SB, AF₄]), which gives the existence of area-minimizing rectifiable currents, as a preliminary regularity result because it provides some initial geometric structure to the area minimizer. Regularity

theory is concerned with showing to what extent the support of such a minimizer T is a smooth manifold. For $n = 1$, this is elementary as $\text{spt } T$ is a finite union of disjoint intervals. For $n = 2$ and $k = 1$, W. Fleming [FW], using a result of E. Reifenberg [R], proved the interior regularity result that $\text{spt } T \sim \Gamma$ is an embedded real analytic minimal surface. Combining ideas from DeGiorgi's regularity theory of oriented frontiers [DG] and from [FW] with a study of minimal hypercones, the work of D. Triscari [TD₁, TD₂], F. J. Almgren, Jr. [A₁], and J. Simons [SJ] extended this complete interior regularity for codimension one to dimensions $n \leq 6$. Moreover, it followed that the possible singularities for $n = 7$ are isolated. After E. Bombieri, E. DeGiorgi, and E. Giusti [BDG] established that singularities may actually occur beginning in dimension 7, H. Federer [F₂] gave an inductive argument showing that the interior singular set of a codimension one area minimizer always has Hausdorff dimension at most $n - 7$. For a minimizer of a smooth parametric elliptic integrand in codimension one, the interior singular set was shown to have \mathcal{H}^n measure zero in [AF₂] and later, by [ASS] and [W₄, 5.1], to have Hausdorff dimension strictly less than $n - 2$. In case the smooth submanifold Γ lies on the boundary of a uniformly convex body, the work of W. K. Allard [AW₁ or AW₂] established the regularity, for all k and n , of an area minimizing T near Γ . This was extended in [H₁] to minimizers of a constant coefficient parametric integrand. The extremality hypothesis on the location of Γ was eliminated in [HS₁] which treated codimension one area minimizers. B. White [W₁] extended the latter work to treat boundaries which may have higher multiplicities. The complete interior and boundary regularity of a codimension 1 area minimizer of dimension ≤ 6 leads, by an argument of [TF], to the finiteness of the number of minimizers having a given boundary [MF₁, 4.3].

Regularity theory of area minimizers in codimension $k \geq 2$ has several serious difficulties. In particular, small area excess no longer implies regularity, as seen, for example, with the varieties $\{(y, z) \in \mathbb{C}^2 : z^2 = \varepsilon y^3\}$, which are singular and area minimizing for every positive ε . Approximation by harmonic functions must include multivalued functions which are not even C^1 and do not provide interior estimates as used in the codimension one theory. In the important work [AF₃], F. J. Almgren attacks these problems by developing several notions: approximation by graphs of multivalued functions, regularity theory for energy-minimizing multivalued functions and their graphs, monotonicity of frequency for such functions, and a higher order "center manifold" approximation theorem. The main result of [AF₃] is that the interior singular set of an area minimizer has Hausdorff dimension at most $n - 2$. Moreover, for $n = 2$, these singularities are isolated in \mathbb{R}^{n+k} by the thesis of S. Chang [C]. The minimality of holomorphic varieties shows the optimality of these results.

Energy minimizing maps with $p = 2$ (or more generally harmonic maps) arise naturally in many problems in geometry and analysis [EL, J₂]. In case there is some restriction on N (e.g., negatively curved or lying in a coordinate neighborhood) there are several interesting early works, e.g., [ES, HR, HKW].

For discussion of these and many other works, we refer to the excellent surveys of S. Hildebrandt [H] and J. Jost [J₁, J₂]. In case $p > 2$ and N lies in a coordinate neighborhood, the work of N. Fusco and J. Hutchinson [FH] and of M. Giaquinta and G. Modica [GM] implies, among other things, the partial $C^{1,\alpha}$ regularity of an energy minimizer, for some $0 < \alpha < 1$. Some special cases have been considered earlier by L. C. Evans [E].

With no restriction on N and $p = 2$, the fundamental work of R. Schoen and K. Uhlenbeck [SU₁] (see also [GG, SU₂, JM, SU₃]) showed that the interior singular set of an energy minimizer has Hausdorff dimension at most $m - 3$. The study of liquid crystals leads to consideration of a more general energy functional with quadratic growth for mappings from 3-dimensional spatial domains to S^2 . For these minimizers, R. Hardt, D. Kinderlehrer, and F. H. Lin [HKL₁] showed that the singular set had 1-dimensional Hausdorff measure zero. For $p > 2$ and more general energy functionals with p -power growth, S. Luckhaus [LS₁, LS₂] established the $C^{1,\alpha}$ regularity of minimizers away from a singular set of dimension $m - p$. In the independent work [HL₂], p -energy minimizers, for any $1 < p < \infty$, were shown to be $C^{1,\alpha}$ regular away from an interior singular set of dimension at most $m - [p] - 1$. For more general functionals with p -power growth, arguments in [HKL₂] lead to the (not necessarily optimal) dimension estimate $m - p - \varepsilon$ for some positive ε . For $p \neq 2$, the highest regularity that can be expected is $C^{1,\alpha}$ for some $0 < \alpha < 1$ (although there may be a partial higher regularity result.) For $p > 2$, the corresponding problem with N replaced by \mathbb{R}^n was first treated in the work of K. Uhlenbeck [U]. To handle all p with $1 < p < \infty$, [HL₂] combines arguments of E. DiBenedetto [DB] and P. Tolksdorf [TP].

5. Asymptotic behavior near a singularity. An important property of an area minimizing current T or a p -minimizing mapping u , as contrasted with minimizers of more general parametric integrands or functionals, is the monotonicity in r of the normalized area in a ball $r^{-n} \text{Area}(T \cap \mathbb{B}_r^{n+k}(a))$ or the normalized energy on a ball $r^{p-m} \int_{\mathbb{B}_r^n(a)} |\nabla u|^p dx$. This leads to the existence of area-minimizing tangent cones and of tangent maps. These are limits of normalized sequences of currents or mappings corresponding to a point a in $\text{spt } T \sim \Gamma$ or in $M \sim \partial M$ and to a sequence of numbers $r_i \downarrow 0$. To use these tangent objects to describe the asymptotic behavior of the minimizer on approach to a , one needs to know that they are, at least, independent of the choice of r_i . While this uniqueness question is open in general, some important cases have been treated. B. White [W₂] verified this for 2-dimensional area-minimizing currents. F. J. Almgren and W. Allard [AA] proved the uniqueness in case one of the tangent cones at a of an area-minimizing (or even stationary) current T is the cone over a (multiplicity one) regular submanifold of S^{n+k-1} , and in case each Jacobi field of this submanifold arises from a 1-parameter family of minimal submanifolds. Another proof is given in [S₃, §6] in a general setting applicable to other problems, in particular, harmonic maps from a 3-dimensional domain to S^2 [GW]. In [AA] and [S₃, §6] T approaches its unique tangent cone as fast

as $c \cdot |x - a|^\alpha$ for some $\alpha > 1$. While the Jacobi field integrability condition is true for a product of two spheres [AA], its validity is unknown for codimension one minimizing cones and can be false for higher codimension. If this condition does fail for some regular minimal cone, D. Adams and L. Simon [AS] construct a stationary T which approaches this cone at a rate *slower* than $|x|^\alpha$ for any $\alpha > 1$, with similar results for harmonic maps. R. Gulliver and B. White [GW] also construct a harmonic map from \mathbb{B}^3 to a real analytic surface N which approaches its unique tangent map at a singular point a at a rate slower than any positive power of $|x - a|$.

In the important paper [S₁], L. Simon eliminates the Jacobi field hypothesis from [AA], and obtains the (in general, slow) decay to a unique tangent cone. Formulated as an asymptotic property of solutions of a certain class of partial differential equations, this work implies, in particular, the uniqueness of tangent cones for 7-dimensional area minimizing rectifiable currents in codimension one and the uniqueness of tangent maps for energy minimizing harmonic maps with 3-dimensional domains. The proof involves a careful study of the approximation by and asymptotics of solutions of the corresponding linearized system. Here, for the application to rectifiable currents, in contrast to harmonic maps, the solution of the nonlinear equation is, by graphical approximation, not defined *a priori* up to the singularity, and one part of the problem is guaranteeing the extendability of such solutions.

6. Behavior under perturbations. For the minimization problems of §1, one may ask about the effect on the singular set caused by smooth perturbations of the boundary data (φ or Γ). Of course, if there is a topological obstruction that is unchanged by the perturbation, there will still be a singularity. In the absence of topological obstruction, singularities may or may not disappear under perturbation of boundary data, as can be seen by varying the coefficients of polynomials defining a complex variety.

For codimension one area minimizers, there are several relevant remarks. The complement in the ambient Euclidean space of the minimizing cone of [BDG], as well as other specific minimizing hypercones C [LB, FK] was shown to be foliated by regular minimal hypersurfaces, each of which, when restricted to a compact set, defines an area-minimizing current. In particular, perturbing the boundary $\Gamma = C \cap \partial\mathbb{B}$ by moving to one of these leaves eliminates the singularity in the minimizer (which is here unique). This behavior generalizes. It is shown in [HS₂] that any area-minimizing cone with singularity only at the vertex gives rise to a unique such minimal foliation. Moreover, the singularities disappear for any small *one-sided* perturbation of a smooth boundary for a minimizing hypersurface with only isolated singularities [HS₂]. A more general sufficient condition for perturbing away singularities is given in [MR]. It should be noted that rectifiable currents may be *one-sided minimizing* [LF₁, W] and that a minimal foliation may exist on only one side of a minimal (not minimizing) hypercone [LF₂]. For two-sided perturbations, on the other hand, there are

many ways [CHS] to make a sufficiently balanced perturbation of $\Gamma = C \cap \partial\mathbb{B}$ to give a minimal hypersurface that is not itself a cone, but is asymptotic to C and has an isolated singularity at 0. Many of these are also area-minimizing [HS₂]. In the absence of a boundary, B. White has recently shown [W₄] the existence of a singular complete minimal hypersurface that is not a cone. He also obtains [W₄] an area-minimizing hypersurface with asymptotic cone C which is not C itself or a leaf of the foliation associated with C . This is in contrast to the result of L. Simon and B. Solomon [SS] that C and its associated leaves exhaust the area-minimizing hypersurfaces asymptotic to C in case $C \cap \partial\mathbb{B}$ is the product of two spheres.

7. Density bounds. A rough measure of the complexity of a singularity at a is the *density*

$$\Theta(u, a) = \lim_{r \downarrow 0} r^{p-m} \int_{\mathbb{B}_r^m(a)} |\nabla u|^p dx \quad \text{for a } p\text{-minimizing map } u,$$

or

$$\Theta(T, a) = \lim_{r \downarrow 0} r^{-n} \text{Area}(T \cap \mathbb{B}_r^{n+k}(a)) \quad \text{for an area-minimizing current } T.$$

They give essentially the energy in the unit ball of any tangent map of u at a or the area in a unit ball of any tangent cone of T at a . Both numbers exist and are upper semicontinuous in a by monotonicity. In particular, they are, for fixed u and T , locally bounded away from the boundary.

In [HKL₂] it is shown that for any p -energy minimizing map u from M into a simply $[p] - 1$ connected N (i.e., $\Pi_0(N) = \Pi_1(N) = \cdots = \Pi_{[p]-1}(N) = 0$) and point $a \in M$, $\Theta(u, a) \leq c$ for some universal bound $c = c(M, N, p)$ not depending on u or a . This upper bound, with $\lim_{r \downarrow 0}$ replaced by $\limsup_{r \downarrow 0}$ and with the same restriction on N , continues to hold for more general functionals including the liquid crystal functional [HKL₁]. For 2-minimizing maps from a 3-dimensional domain M to \mathbb{S}^2 , the work [BCL] of H. Brezis, J. P. Coron, and E. Lieb says much more. They show that the only nonconstant minimizing tangent maps are specifically the degree one maps $g(x/|x|)$ corresponding to rotations $g \in \mathcal{O}(3)$. In general, homogeneous-degree-0 harmonic maps from \mathbb{B}^3 to \mathbb{S}^2 are classified by rational maps: $\mathbb{C} \rightarrow \mathbb{C}$ [J₂, 1.5]. The numerical work [CHKLL] indicates how energy minimization leads to a splitting in the singularity of the homogeneous extension of the higher degree maps. [BCL] also contains a precise description of limiting configurations for a prescribed singularity problem as well as an energy lower bound for maps $u: M \rightarrow \mathbb{S}^2$ with $u|_{\partial M}$ having constant sign Jacobian.

Among codimension one minimizers of area (or other parametric integrands), such a universal density bound holds for oriented frontiers. But in higher codimensions, the minimality of holomorphic varieties of high degree shows the impossibility of such a universal bound. For the somewhat different problem of minimizing chains modulo ν , F. Morgan [MF₂] has obtained such a bound for all dimensions and codimensions.

8. Questions. (1) In partial regularity theory, there are many open problems. The optimal estimate of the singular set is, in general, unknown for functionals other than p -energy or area. Even for each of these, the structure of the singular set is understood only in the lowest dimension where it may occur. What is the regularity of the singular set? Is it of finite measure in the appropriate dimension?

(2) What is the asymptotic behavior of the minimizer near a nonisolated singularity? Is the tangent cone or tangent map unique? The only work in this direction is L. Simon's study [S₄] of the asymptotics of nonplanar complete minimal graphs. Here the tangent cone at infinity is a cylinder containing a line of singularities.

(3) In the absence of topological obstructions, do singularities in minimizers disappear under generic perturbation of the boundary data. B. White [W₁] showed this in the somewhat different context of surfaces that minimize area among oriented and unoriented surfaces (such minimizers are immersed surfaces [R, AF₁]).

(4) What is the regularity of nonminimizing stationary rectifiable currents or mappings? Applications of Allard's general regularity theorem [AW₂] have been limited by multiplicity one hypothesis. Regularity results for stationary harmonic maps with no restriction on the target manifold are only known for $m = 2$ [SR]. See the discussion of J. Jost [B, 4.11]. G. Liao [LG] showed that an isolated singularity of a small energy stationary harmonic map is removable.

(5) For other interesting related questions see the extensive list [B].

REFERENCES

- [AS] D. Adams and L. Simon, *Asymptotic behavior near isolated singular points for geometric variational problems*, Proc. Centre Math. Anal. Austral. Nat. Univ., vol. 10, Austral. Nat. Univ., Canberra, 1985.
- [AW₁] W. K. Allard, *On boundary regularity for the Plateau problem*, Thesis, Brown University, 1968.
- [AW₂] —, *On the first variation of a varifold*, Ann. of Math. (2) **95** (1972), 417–491.
- [AW₃] —, *On the first variation of a varifold: boundary behavior*, Ann. of Math. (2) **101** (1975), 418–446.
- [AF₁] F. J. Almgren, Jr., *Some interior regularity theorems for minimal surfaces and an extension of Bernstein's theorem*, Ann. of Math. (2) **84** (1966), 277–292.
- [AF₂] —, *Existence and regularity almost everywhere of solutions to elliptic variational problems among surfaces of varying topological type and singularity structure*, Ann. of Math. **87** (1968), 321–391.
- [AF₃] —, *Q valued functions minimizing Dirichlet's integral and the regularity of area minimizing rectifiable currents up to codimension 2*, Preprint.
- [AF₄] —, *Deformations and multiple-valued functions*, Proc. Sympos. Pure Math., Vol. 44, Amer. Math. Soc., Providence, R. I., 1980, pp. 29–130.
- [AA] W. K. Allard and F. J. Almgren, Jr., *On the radial behavior of minimal surfaces and the uniqueness of their tangent cones*, Ann. of Math. (2) **113** (1981), 215–265.
- [ASS] F. J. Almgren, Jr., R. Schoen, and L. Simon, *Regularity and singularity estimates on hypersurfaces minimizing parametric elliptic variational integrals*, Acta Math. **139** (1977), 217–265.
- [BCL] H. Brezis, J.-M. Coron, and E. Lieb, *Estimations d'énergie pour des applications de \mathbb{R}^3 à valeurs dans S^2* , C. R. Acad. Sci. **303** (1986), 207–210.

- [B] J. Brothers, *Some open problems in geometric measure and its applications suggested by participants of the 1984 AMS Summer Institute*, Proc. Sympos. Pure Math., Vol. 44, Amer. Math. Soc., Providence, R. I., 1980, pp. 441–464.
- [BDG] E. Bombieri, E. De Giorgi, and E. Giusti, *Minimal cones and the Bernstein problem*, Invent. Math. **7** (1969), 243–268.
- [C] S. Chang, Thesis, Princeton University, 1986.
- [CHS] L. Caffarelli, R. Hardt, and L. Simon, *Minimal surfaces with isolated singularities*, Manuscripta Math. **48** (1984), 1–18.
- [CHKLL] R. Cohen, R. Hardt, D. Kinderlehrer, S. Y. Lin, and M. Luskin, *Minimum energy configurations for liquid crystals: computational results*, Theory and Applications of Liquid Crystals, IMA vol., Springer, New York (to appear).
- [DB] E. DiBenedetto, $C^{1+\alpha}$ local regularity of weak solutions of degenerate elliptic equations, Nonlinear Anal. **7** (1983), no. 8, 827–850.
- [DG] E. Di Giorgi, *Frontiere orientate di misura minima*, Seminario Mat. Scuola Norm. Sup. Pisa, 1961.
- [E] L. C. Evans, *Quasi convexity and partial regularity in the calculus of variations*, University of Maryland, Preprint, 1984.
- [EL] J. Eels and L. Lemaire, *A report on harmonic maps*, Bull. London Math. Soc. **10** (1978), 1–68.
- [ES] J. Eels and J. H. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math. **86** (1964), 109–160.
- [F₁] H. Federer, *Geometric measure theory*, Springer-Verlag, Berlin, Heidelberg, and New York, 1969.
- [F₂] —, *The singular sets of area minimizing rectifiable currents with codimension one and of area minimizing flat chains modulo two with arbitrary codimension*, Bull. Amer. Math. Soc. **76** (1970), 767–771.
- [F₃] —, *Colloquium lectures on geometric measure theory*, Bull. Amer. Math. Soc. **84** (1978), 291–338.
- [FF] H. Federer and W. Fleming, *Normal and integral currents*, Ann. Math. **72** (1960), 480–520.
- [FK] D. Ferus and H. Karcher, *Non-rotational minimal spheres and minimizing cones*, Comment. Math. Helv. **60** (1985), 247–269.
- [FW] W. Fleming, *On the oriented Plateau problem*, Rend. Circ. Mat. Palermo (2) **11** (1962), 1–22.
- [FH] N. Fusco and J. Hutchinson, *Partial regularity for minimizers of certain functionals having non quadratic growth*, Preprint, CMA, Canberra, 1985.
- [GG] M. Giaquinta and E. Giusti, *The singular set of the minima of certain quadratic functionals*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **11** (1984), no. 1, 45–55.
- [GM] M. Giaquinta and G. Modica, *Remarks on the regularity of the minimizers of certain degenerate functionals*, Preprint.
- [GW] R. Gulliver and B. White, *The rate of convergence of a harmonic map at a singular point*, Preprint.
- [HR] R. Hamilton, *Harmonic maps of manifolds with boundary*, Lecture Notes in Math., Vol. 471, Springer-Verlag, Berlin and New York, 1975.
- [H₁] R. Hardt, *On boundary regularity for integral currents or flat chains modulo 2 minimizing the integral of an elliptic integrand*, Comm. Partial Differential Equations **2** (1977), 1163–1232.
- [HKL₁] R. Hardt, D. Kinderlehrer, and F. H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Comm. Math. Phys. **105** (1986), 547–570.
- [HKL₂] —, *Energy bounds for minimizing mappings* (in preparation).
- [HL₁] R. Hardt and F. H. Lin, *A remark on H^1 mappings*, Manuscripta Math. (to appear).
- [HL₂] —, *Mappings minimizing the L^p norm of the gradient*, Comm. Pure Appl. Math. (to appear).
- [HS₁] R. Hardt and L. Simon, *Boundary regularity and embedded solutions for the oriented Plateau problem*, Ann. of Math. (2) **110** (1979), 439–486.

- [HS₂] —, *Area minimizing hypersurfaces with isolated singularities*, J. Reine Angew. Math. **362** (1985), 102–129.
- [H] S. Hildebrandt, *Harmonic mappings of Riemannian manifolds*, Lecture Notes in Math., Vol. 1161, Springer-Verlag, Berlin and New York, 1985.
- [HKW] S. Hildebrandt, H. Kaul, and K. Widman, *Harmonic mappings into Riemannian manifolds with non-positive sectional curvature*, Math. Scand. **37** (1975), 257–263.
- [J₁] J. Jost, *Harmonic mappings between Riemannian manifolds*, Proc. Centre Math. Anal. Austral. Nat. Univ., Austral. Nat. Univ., Canberra, 1984.
- [J₂] —, *Lectures on harmonic maps*, Lecture Notes in Math., Vol. 1161, Springer-Verlag, Berlin and New York, 1985.
- [JM] J. Jost and M. Meier, *Boundary regularity for minima of certain quadratic functionals*, Math. Ann. **262** (1983), 549–561.
- [KJF] A. Kufner, O. John, and S. Fučík, *Function spaces*, Noordhoff, Leyden, 1977.
- [LB] H. B. Lawson, *The equivariant Plateau problem and interior regularity*, Trans. Amer. Math. Soc. **173** (1973), 231–249.
- [LG] G. Liao, *A regularity theorem for harmonic maps with small energy*, J. Differential Geom. **22** (1985), 233–241.
- [LF₁] F. H. Lin, *Plateau's problem for H -convex curve*, Manuscripta Math. (to appear).
- [LF₂] —, *Minimality and stability of minimal hypersurfaces in \mathbb{R}^n* , Preprint.
- [LS₁] S. Luckhaus, *Partial Hölder continuity of energy minimizing p -harmonic maps between Riemannian manifolds*, Preprint, CMA, Canberra, 1986.
- [LS₂] —, *$C^{1,\varepsilon}$ -regularity for energy minimizing Hölder continuous p -harmonic maps between Riemannian manifolds*, Preprint, CMA, Canberra, 1986.
- [MR] R. McIntosh, *Perturbing away singularities from area-minimizing hypercones*, Preprint, IAS, Austral. Nat. Univ., 1985.
- [MF₁] F. Morgan, *On finiteness of the number of stable minimal hypersurfaces with a fixed boundary*, Preprint.
- [MF₂] —, *Mass bounds and Bernstein theorems for area-minimizing flat chains modulo ν* , Preprint.
- [R] R. E. Reiffenberg, *Solution of the Plateau problem for m -dimensional surfaces of varying topological type*, Acta Math. **104** (1960), 1–92.
- [SU₁] R. Schoen and K. Uhlenbeck, *A regularity theory for harmonic maps*, J. Differential Geom. **17** (1982), 307–335.
- [SU₂] —, *Boundary regularity and the Dirichlet problem of harmonic maps*, J. Differential Geom. **18** (1983), 253–268.
- [SU₃] —, *Regularity of minimizing harmonic maps into the sphere*, Invent. Math. **16** (1984), 89–100.
- [SR] R. Schoen, *Analytic aspects of the harmonic map problem*, Preprint.
- [S₁] L. Simon, *Asymptotics for a class of nonlinear evolution equations*, Ann. of Math. (2) **118** (1983), 525–571.
- [S₂] —, *Lectures on geometric measure theory*, Proc. Centre Math. Anal. Austral. Nat. Univ., Vol. 3, Austral. Nat. Univ., Canberra, 1984.
- [S₃] —, *Isolated singularities for extrema of geometric variational problems*, Lecture Notes in Math., Vol. 1161, Springer-Verlag, Berlin and New York, 1985.
- [S₄] —, *Entire solutions for the minimal surface equation* (in preparation).
- [SJ] J. Simons, *Minimal varieties in Riemannian manifolds*, Ann. of Math. (2) **88** (1968), 62–105.
- [SS] L. Simon and B. Solomon, *Minimal hypersurfaces asymptotic to a quadratic cone in \mathbb{R}^{n+1}* , Preprint, CMA, Canberra, 1985.
- [SB] B. Solomon, *A new proof of the closure theorem for integral currents*, Indiana Univ. Math. J. (3) **33** (1984), 393–418.
- [TP] P. Tolksdorf, *Everywhere regularity for some quasi-linear systems with lack of ellipticity*, Ann. Mat. Pura Appl. **134** (1983), 241–266.
- [TF] F. Tomi, *On the finite solvability of Plateau's problem*, Lecture Notes in Math., Vol. 597, Springer-Verlag, Berlin and New York, 1977, pp. 679–695.
- [TD₁] D. Triscari, Ann. Scuola Norm. Sup. (3) **17** (1963), 349–371, 387–399.

[TD₂] —, *Sulle singularita delle frontiere orientate di misura minima nello spazio euclideo a 4 dimensione*, Matematiche (Catania) **18** (1963), 139–163.

[U] K. Uhlenbeck, *Regularity of nonlinear elliptic systems*, Acta Math. **138** (1977), 219–240.

[W₁] B. White, *Singularity structure and generic regularity of 2-dimensional area minimizing surfaces*, Thesis, Princeton Univ., 1981.

[W₂] —, *Tangent cones to 2-dimensional area minimizing currents are unique*, Duke Math. J. **50** (1983), 143–160.

[W₃] —, *Regularity of area-minimizing hypersurface at boundaries with multiplicity*, Ann. of Math. Studies, Vol. 103, Princeton Univ. Press, Princeton, N. J., 1983.

[W₄] —, *A regularity theorem for minimizing hypersurfaces mod p* , Proc. Sympos. Pure Math., Vol. 44, Amer. Math. Soc., Providence, R. I., 1980, pp. 413–428.

[W₅] —, *New applications of mapping degrees to minimal surface theory*, Preprint, CMA, Canberra, 1986.

SCHOOL OF MATHEMATICS, UNIVERSITY OF MINNESOTA, MINNEAPOLIS, MINNESOTA 55455, USA

Recent Progress on the Geometry of Surfaces in \mathbb{R}^3 and on the Use of Computer Graphics as a Research Tool

W. H. MEEKS III

The progress in the last decade in the geometry of surfaces in three-space has been nothing less than spectacular. The major part of this research has been concerned with the theory of constant mean curvature surfaces and, in particular, with the theory of minimal surfaces. The intent of the author's lecture at the International Congress of Mathematicians 1986 was to give a brief survey of the most recent results on the geometry of surfaces, with an emphasis on applications of computer graphics. The author apologizes in advance for the many omissions and necessary restriction of topics in this present survey. The author would like to thank Jim Hoffman, whose expertise in computer graphics motivated much of the material of my talk and who made the beautiful surface pictures which I presented at the Congress in the form of slides.

1. Recent progress on the classical theory of minimal surfaces. The most important development in the classical theory of minimal surfaces in the past three years has been in that part of the theory dealing with properly embedded examples of finite type. A surface has *finite type* if it is homeomorphic to a closed surface with a finite number of points removed. The first advance in the development of a theory to deal with these special minimal surfaces was a series of existence theorems by Hoffman and Meeks [18, 19, 20, 21] and by Callahan, Hoffman, and Meeks [4]. Many of the techniques used in these existence theorems are quite general and of intrinsic interest. Also, explicit analytical representations of these surfaces, together with the expertise of computer graphics programmer J. Hoffman, has given the general public the rare opportunity to see these beautiful surfaces in popular science magazines [17, 24, 40]. Hoffman and Meeks [20] have subsequently developed a general theory to deal with the new properly embedded examples. Recently, several further discoveries in the

Research was partially supported by National Science Foundation Grant DMS-8611574 and Department of Energy Grant FG02-86ER25015.

global theory of embedded minimal surfaces have appeared. These results give topological, conformal, and geometrical restrictions on the behavior of properly embedded minimal surfaces, and lead to many fascinating conjectures on their global structure. We will discuss these later in the paper.

Until 1984 the only known examples of properly embedded minimal surfaces of finite type were the plane, the catenoid, and the helicoid. Under the additional hypothesis that the surface had finite *total Gaussian curvature* (which is the integral of the absolute value of the Gaussian curvature function on the surface) Jorge and Meeks [23] had shown that there were no examples which were homeomorphic to S^2 punctured in 3, 4, or 5 points. Later R. Schoen [38] proved that if a complete embedded minimal surface has finite total curvature and is homeomorphic to a closed surface punctured in 2 points, then it is a catenoid. These negative results supported the generally-believed conjecture that the only examples were the plane, the catenoid, and the helicoid.

Around 1981 C. Costa [7], in his thesis at IMPA, defined a properly immersed minimal surface homeomorphic to a torus punctured in 3 points, which he showed to be embedded outside of a compact ball. In the spring of 1984, using a graphic system developed by J. Hoffman, J. Hoffman and D. Hoffman made computer graphics pictures of Costa's surface and observed that it seemed to be embedded. Shortly after this discovery, Hoffman and Meeks [19] gave a rigorous proof that Costa's surface is embedded. Using symmetry, classical Riemann surface theory, and the analytical Weierstrass representation of a minimal surface, Hoffman and Meeks gave an explicit analytic representation of the infinite family of minimal surfaces M_k described in the next theorem [19, 20, 21]. Shortly after this, they found an independent and more abstract existence proof of these examples. This second proof gives a more satisfactory answer to the deep question "Why do these new examples of properly embedded minimal surfaces of finite type exist?"

THEOREM 1. *For each positive integer k , there exists a properly embedded minimal surface M_k which is homeomorphic to a closed surface with 3 points removed. Furthermore, M_k has the following properties:*

- (1) M_k has finite total curvature $-4\pi(k+2)$.
- (2) The symmetry group of M_k is the usual Z_2 -extension in $O(3)$ of a dihedral group $D(k+1)$ in $SO(3)$.
- (3) M_k intersects the xy -plane P in $k+1$ straight lines making equal angles at the origin.
- (4) If P' is a plane parallel to P , $P' \neq P$, then $P' \cap M_k$ is a simple closed curve.
- (5) M_k has maximal symmetry, in the sense that any homeomorphic properly embedded minimal surface N_k of finite total curvature having a symmetry group as large as M_k 's must be identical to M_k after a homothety and a rigid motion of \mathbb{R}^3 .
- (6) After a translation \tilde{M}_k of M_k , the surfaces \tilde{M}_k converge smoothly as subsets of \mathbb{R}^3 to Sherk's second surface.

(7) Each surface M_k is part of a 1-parameter family of minimal surfaces $M_k(t)$. For $t \in \mathbf{R}^+$ the surface $M_1(t)$ is conformally equivalent to \mathbf{C}/L_t , where $L_t = \{n + mt\sqrt{-1}, m, n \in \mathbf{Z}\}$, punctured in the 3 half-lattice points.

Recently Callahan, Hoffman, and Meeks [4] have shown that there exist properly embedded minimal surfaces of finite total curvature homeomorphic to closed surfaces of even genus with 4 points removed. Their technique relies on the computer not only to depict the surface graphically but also to solve for constants needed in the actual analytic construction of the surfaces.

The next four theorems give some nontrivial information about the global geometry and topology of certain minimal surfaces in \mathbf{R}^3 . The proofs of these theorems are quite geometric and frequently use recent results on the geometry of stable minimal surfaces.

THEOREM 2 (FISCHER-COBRIE [11]). *If $f: M \rightarrow \mathbf{R}^3$ is a complete minimal immersion of an orientable surface, then M has finite index (with respect to the stability operator) if and only if M has finite total curvature.*

THE STRONG HALF-SPACE THEOREM (HOFFMAN AND MEEKS [20]). *If M_1 and M_2 are two properly immersed minimal surfaces which are not parallel planes, then M_1 and M_2 intersect at some point.*

THE ISOMETRY THEOREM (CHOI, MEEKS, AND WHITE [6]). *If M is a properly embedded minimal surface in \mathbf{R}^3 and contains a simple closed curve which separates M into 2 noncompact components, then every intrinsic isometry of M extends to an ambient isometry of \mathbf{R}^3 .*

THE ANNULAR END THEOREM (HOFFMAN AND MEEKS [20]). *Let M be a properly embedded minimal surface in \mathbf{R}^3 . An annular end of M is a proper annulus contained in M which is homeomorphic to $S^1 \times [0, 1)$. Then M can have at most 2 pairwise disjoint annular ends with infinite total curvature. In particular, M cannot be conformally diffeomorphic to a Riemann surface with 3 compact pairwise disjoint disks removed.*

THE TOPOLOGICAL UNIQUENESS THEOREM (MEEKS AND YAU [34]). *Suppose M_1 and M_2 are two properly embedded homeomorphic minimal surfaces in \mathbf{R}^3 which have finite type. Then there exists a homeomorphism $h: \mathbf{R}^3 \rightarrow \mathbf{R}^3$ such that $h(M_1) = M_2$.*

The above theorems lead naturally to the following conjectures.

CONJECTURE 1. *For every positive integer k , there exists a minimal surface of finite type homeomorphic to a closed surface with k points removed.*

CONJECTURE 2. *A properly embedded minimal surface of finite type and infinite total curvature must be a helicoid.*

CONJECTURE 3. *An intrinsic isometry on a properly embedded minimal surface in \mathbf{R}^3 extends to an ambient isometry.*

Recently there has been a lot of interest in doubly and triply periodic minimal surfaces. The reason for this interest is that such surfaces appear as, or are related to, surface interfaces arising in chemistry, polymer science, the theory of liquid crystals, and in the study of Fermi surfaces in solid state physics. Below are some recent results on the existence and geometry of these special surfaces.

THEOREM (MEEKS-ROSENBERG [33]). *Suppose M is a doubly periodic embedded minimal surface in \mathbf{R}^3 with translational symmetry group given by $L = \{nv_1 + mv_2 | n, m \in \mathbf{Z}\}$. If the quotient surface $\overline{M} = M/L$ has finite topology, then \overline{M} has finite total Gaussian curvature.*

THE ABEL-GAUSS-BONNET THEOREM (MEEKS [30]). *Suppose $f: M_g \rightarrow T^3 = \mathbf{R}^3/L$ is a minimal immersion of a closed surface of genus g into a flat 3-torus with lattice L . Then:*

(1) *The Gauss map $g: M_g \rightarrow S^2 = \mathbf{C} \cup \{\infty\}$ is a meromorphic function when M_g is considered to be a Riemann surface.*

(2) *Consider T^3 to be an abelian group and define $\hat{f}: S^2 \rightarrow T^3$ by $\hat{f}(v) = \sum_{x \in g^{-1}(v)} x$. Then \hat{f} is constant.*

(3) *If $g = 3$ and M_3 is embedded, then M_3 is invariant under an inversion through any zero of Gaussian curvature.*

(4) *If M_g is embedded and $g > 1$, then M_g disconnects T^3 into two 1-handle bodies. In particular, by part (3), if $g = 3$, then M_3 disconnects T^3 into isometric pieces.*

Parts (1) and (3) of the above theorem were also observed by Nagano and Smyth [38] at about the same time as Meeks. Actually, part (1) is rather obvious and had been observed earlier by others.

THE FUNDAMENTAL EXISTENCE THEOREM (MEEKS [30, 32]). *Let T^3 be a flat 3-torus. Then there exists a sequence of embedded minimal surfaces N_1, \dots, N_i, \dots of genus 3 such that $\text{Area}(N_k) \geq k$.*

Part of the proof of the Fundamental Existence Theorem depends on the following generalization of the Jacobi embedding theorem for Riemann surfaces.

THE REGULARITY OF THE ALBANESE MAP (MEEKS [31]). *Suppose M is a closed nonorientable Riemannian surface and let h_1, \dots, h_k be a basis for the harmonic 1-forms on M . Let $L = \{\int_\gamma (h_1, \dots, h_k) | \gamma \in H_1(M)\}$. The Albanese map for M is defined to be*

$$A(p) = \int_{p_0}^p (H_1, \dots, H_k): M \rightarrow A(M) \equiv \mathbf{R}^k/L.$$

Then the following statements are equivalent:

- (1) *$A: M \rightarrow A(M)$ is not one-to-one.*
- (2) *$A: M \rightarrow A(M)$ is not an immersion.*
- (3) *There is a curve C on M such that $A|_C$ is constant.*
- (4) *M has a special "conformal" structure.*

A nonorientable “Riemann” surface is called *special* if there is a sequence

$$X = \{p_1, \bar{p}_1, \dots, p_k, \bar{p}_k\}$$

of pairs of distinct conjugate points in \mathbf{C} such that if \tilde{M} denotes the 2-sheeted branched cover of $S^2 = \mathbf{C} \cup \{\infty\}$ branched over X , and $\sigma: \tilde{M} \rightarrow \tilde{M}$ denotes the lift of complex conjugation on $S^2 = \mathbf{C} \cup \{\infty\}$ which acts freely on \tilde{M} , then M is “conformally” equivalent to $\tilde{M}/\{\text{id}, \sigma\}$.

2. Recent progress on the geometry of nonzero constant mean curvature surfaces. If the mean curvature of a surface is constant but not zero, then the surface is frequently called a *soap bubble surface*. The reason for this name is that small pieces of the surface occur as soap films on wires where the air pressure on one side differs from the air pressure on the other side of the surface. If M is a closed embedded surface of constant mean curvature, then it follows from Alexandrov [2] that M must be a round sphere. He proved this theorem with a technique that shows that any closed embedded hypersurface in \mathbf{R}^n which has constant mean curvature is invariant under reflection in a large number of hyperplanes. The technique used in his proof is now known as the Alexandrov reflection principle. This principle is useful in the consideration of other, not necessarily closed, embedded surfaces of constant mean curvature. However, this technique does not work well for closed immersed surfaces of constant mean curvature.

In 1984 H. Wente [46] showed that there exists an infinite number of different immersed tori of constant mean curvature. Recently, U. Abresch [1] has found a rather explicit form representation for some constant mean curvature immersed tori in terms of elliptic functions. By carefully looking at computer graphics images which he created, he observed that the examples of Wente appeared to be foliated by principal curves which are contained in planes. This observation led to a simplification in the equations and to Abresch’s simpler construction. J. Spruck [44] has shown that the examples of Abresch and the examples of Wente are the same. Spruck has also shown that there are other new examples of constant mean curvature tori which are different from the original examples of Wente. Wente [47] has also produced similar new examples.

Recall that a surface has *finite type* if it is homeomorphic to a closed surface punctured in a finite number of points. As discussed in the previous section, there are many properly embedded minimal surfaces of finite type. However, the only known examples of properly embedded nonzero constant mean curvature surfaces of finite type are the sphere, the infinite cylinder, and the periodic Delaunay [8] surfaces of revolution. On the other hand, it seems likely that many new examples will soon be forthcoming.

Very recently, Meeks [29] has proven the following theorem concerning the topology and geometry of surfaces of nonzero constant mean curvature.

THEOREM. *Suppose M is a properly embedded surface in \mathbf{R}^3 with nonzero constant mean curvature. Then:*

- (1) *M is not homeomorphic to a closed surface with a single point removed.*
- (2) *If M is homeomorphic to a closed surface with 2 points removed, then M stays a bounded distance from some straight line.*
- (3) *If M is homeomorphic to a closed surface in 3 points, then M stays a bounded distance from a plane.*
- (4) *If A is a proper annular end of M which is homeomorphic to $S^1 \times [0, 1)$, then A stays a bounded distance from a ray.*

This theorem is proved by studying stable minimal surfaces in one of the complements of the surface in \mathbf{R}^3 . The curvature estimates for stable minimal surfaces given by R. Schoen [42] play an essential role in the analysis of this problem. A first step in proving the above theorem is the following rather surprising proposition, which describes the geometry of properly embedded annuli of mean curvature bounded away from zero.

PROPOSITION. *Let $A = S^1 \times [0, 1)$ and $f: A \rightarrow \mathbf{R}^3$ be a proper embedding. Suppose that $f(A)$ has mean curvature greater than 1 and P_1 is a plane such that $P_1 \cap (f(A))$ contains a noncompact connected component. Then there exists a universal constant C independent of f such that if P_2 is a plane which is transverse to $f(A)$ and the distance from P_1 to P_2 is greater than C , then every component of $P_2 \cap (f(A))$ consists of simple closed curves.*

In a somewhat different direction, B. Palmer [34] and A. Silveira [36] proved that a complete stable surface of constant mean curvature is either a round sphere or a flat plane. Intuitively, stability means that these surfaces are physically realizable as soap bubbles. Their result generalizes the earlier, similar theorem for minimal surfaces by do Carmo and Peng [5] and Fischer-Cobrie and Schoen [12]. A simple consequence of this result is that a foliation of an open set of \mathbf{R}^3 by complete surfaces of a fixed constant mean curvature is a foliation by parallel planes [3]. This observation, together with the general estimates used in the proof of the earlier stated theorem of Meeks, easily imply that the only possible foliation of \mathbf{R}^3 by surfaces of possibly varying, constant mean curvature is one by parallel planes [29].

A general question that arises from H. Wente's constant mean curvature tori is whether or not there exist immersed surfaces of constant mean curvature which have genus greater than 1. A theorem of H. Hopf [22] shows that any immersed sphere of a constant mean curvature is round.

A second general question concerns the existence of properly embedded non-zero constant mean curvature surfaces of finite type. We conjecture that if such a surface is homeomorphic to a closed surface with 2 points removed, then the surface is one of the surfaces of revolution found by Delaunay [8]. We also conjecture that if M is homeomorphic to a closed surface punctured in at least 3 points, then there exists a proper embedding $f: M \rightarrow \mathbf{R}^3$ with constant mean curvature 1.

A third general question is to what extent the Alexandrov reflection principle can be applied to prove symmetry for general properly embedded surfaces of constant mean curvature. We conjecture that a properly embedded surface of constant mean curvature which is contained in a half-space must be invariant under reflection in a plane parallel to the boundary of the half-space. The validity of this conjecture together with the earlier-stated theorem that a finite type example with 2 ends stays a bounded distance from a line easily implies that the surface must be a surface of revolution, and hence, topologically a sphere punctured in 2 points. Recall from §1 [20] that a proper minimal surface which is contained in a half-space must be a flat plane: this shows that the conjecture is true for zero mean curvature.

3. Recent progress on the theory of isotopy tight surfaces. The most important recent result on the geometry of isotopy tight surfaces is a theorem of Kuiper and Meeks [26] which gives a complete generalization of the Fary-Fenchel-Fox-Milnor theorem for knots to the case of knotted tori. An embedded closed surface M in \mathbf{R}^3 is called *isotopy tight* if the *total Gaussian curvature* $K(M)$ of M , which is the integral of the absolute value of the Gaussian curvature over M , is less than or equal to $K(\overline{M})$ where \overline{M} is any surface which is ambiently isotopic to M .

In various papers, Fary [9], Fenchel [10], Fox [13], and Milnor [31] proved results on the total curvature of a knot in \mathbf{R}^3 . A *knot* is a smooth simple closed curve in space. A knot γ is called an *unknot* if it is isotopic to a round circle. The *total curvature* $K(\gamma)$ of γ is the integral of the curvature of γ . The *bridge number* $B(\gamma)$ of γ is the minimum number of relative maxima that an orthogonal projection of γ onto a line must have. For example, the unit circle γ in the xy -plane has bridge number $B(\gamma) = 1$ since the x -coordinate restricted to γ has exactly one local maximum. If $[\gamma]$ denotes the collection of knots isotopic to γ , then define $K[\gamma]$ to be the infimum of the total curvature of elements in $[\gamma]$.

The combined efforts of Fary-Fenchel-Fox-Milnor led to the following theorem.

THE FARY-FENCHEL-FOX-MILNOR THEOREM. *If γ is a smooth knot, then*

- (1) $K(\gamma) \geq K[\gamma] = 2\pi \cdot B[\gamma] \geq 2\pi$;
- (2) $K(\gamma) = K[\gamma]$ *only for convex plane curves.*

As with curves, one can also define the total curvature of the isotopy class of a surface. Let M be a closed surface embedded in \mathbf{R}^3 and let $[M]$ denote the collection of surfaces which are isotopic to M . Let $K[M]$ be the infimum of the total curvature of surfaces in $[M]$. If T is an embedded torus in \mathbf{R}^3 , then one can consider T as a torus in the three-sphere S^3 by adjoining a point at infinity in \mathbf{R}^3 . A theorem of Alexander [14] shows that one component X of $S^3 \setminus T$ is a 1-handle body homeomorphic to the product of an open disk with a circle. We can choose a *core curve* γ for T , which is a curve γ such that the boundary of a small ε -neighborhood of γ is isotopic in S^3 to T . We can choose γ to be disjoint

from the point at infinity in S^3 . T is said to be *unknotted* if it has an unknotted core curve. The next theorem appears in [26].

THE GENERALIZED FARY-FENCHEL-FOX-MILNOR THEOREM. *If T is a smooth embedded torus in three-space, and γ is a core curve for T , then*

- (1) $K(T) \geq K[T] = 8\pi \cdot B[\gamma] \geq 8\pi$;
- (2) $K(T) = K[T]$ implies $K(T) = 8\pi$ and that T is unknotted.

The inequality $K(T) \geq 16\pi$ for a knotted torus was first proven by Langevin and Rosenberg [27]. The partial result that $K(T) \leq 16\pi$ implies T is unknotted is contained in the earlier paper [25] where Kuiper and Meeks developed a general theory for understanding the geometry and topology of isotopy tight surfaces. A surface M of genus g is called *tight* if $K(M) = 4\pi(g + 1)$, which is the minimum possible value for an immersed surface of genus g . One consequence of the Generalized Fary-Fenchel-Fox-Milnor Theorem is that an isotopy tight torus is, in fact, tight. However, there are surfaces of genus $g > 2$ which are isotopy tight. We conjecture that an isotopy tight surface of genus 2 is tight.

The Generalized Fary-Fenchel-Fox-Milnor Theorem is proved through geometrical and topological arguments and depends, in an essential way, on the special geometry of isotopy tight surfaces. We shall now give a brief description of their geometry (see [25] for a more complete description). If M is an isotopy tight surface, then it can be expressed as a disjoint collection C of connected subsets $C = \{P_1, P_2, \dots, P_k, N_1, \dots, N_m\}$. The compact components P_i are subsets of smooth convex spheres S_i where $S_i \setminus P_i$ consists of convex planar disk components. The open subsets N_i have nonnegative Gaussian curvature and have the geometric property that if x is a boundary point of N_i contained on a sphere S_j , then a neighborhood of x on N_i is contained in the interior of the open convex ball with boundary S_i . In other words, the subsurfaces N_i go into the convex balls near their boundary curves.

REFERENCES

1. U. Abresch, *Constant mean curvature tori in terms of elliptic functions*, Preprint.
2. A. D. Alexandrov, *Uniqueness theorems for surfaces in the large*. I, Vestnik Leningrad Univ. Math. **11** (1956), 5–17. (Russian)
3. L. Barbosa, J. Gomes, and A. Silveira, personal communication.
4. M. Callahan, D. Hoffman, and W. Meeks, *Embedded minimal surfaces with 4 ends*, Preprint.
5. M. do Carmo and C. K. Peng, *Stable minimal surfaces in \mathbf{R}^3 are planes*, Bull. Amer. Math. Soc. (N.S.) **1** (1979), 903–906.
6. H. Choi, W. Meeks, and B. White, *The isometry theorem for properly embedded minimal surfaces*, Preprint.
7. C. Costa, *Imersões mínimas completas em \mathbf{R}^3 de gênero um e curvatura total finita*, Doctoral Thesis, IMPA, Rio de Janeiro, Brasil, 1982; *Example of a complete minimal immersion in \mathbf{R}^3 of genus one and three embedded ends*, Bull. Soc. Bras. Mat. **15** (1984), 47–54.
8. G. Darboux, *Leçons sur la théorie générale des surfaces*, Première Partie, Gauthier-Villars, Paris, 1941.
9. I. Fary, *Sur la courbure totale d'une courbe gauche faisant un noeud*, Bull. Soc. Math. France **77** (1949), 128–138.

10. W. Fenchel, *Über die Krümmung und Windung geschlossener Raumkurven*, Math. Ann. **101** (1929), 228–253.
11. D. Fischer-Colbrie, *On complete minimal surfaces with finite Morse index in 3-manifolds*, Invent. Math. **82** (1985), 121–132.
12. D. Fischer-Colbrie and R. Schoen, *The structure of complete stable minimal surfaces in 3-manifolds of nonnegative scalar curvature*, Comm. Pure Appl. Math. **33** (1980), 199–211.
13. R. H. Fox, *On the total curvature of some tame knots*, Ann. of Math. **52** (1950), 258–261.
14. M. Greenberg and J. Harper, *Algebraic topology: a first course*, Benjamin-Cummings, Reading, Mass., 1981.
15. D. Hoffman, *The discovery of new embedded minimal surfaces: elliptic functions; symmetry; computer graphics*, Proceedings of the Berlin Conference on Global Differential Geometry (Berlin, June 1984).
16. —, *The construction of families of embedded minimal surfaces*, Proceedings of Variational Methods for Free Surface Interfaces, Springer-Verlag, 1987, pp. 25–36.
17. —, *The computer-aided discovery of new embedded minimal surfaces*, Mathematics Intelligencer (to appear).
18. D. Hoffman and W. Meeks III, *Complete embedded minimal surfaces of finite total curvature*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), 134–136.
19. —, *A complete embedded minimal surface in \mathbf{R}^3 with genus one and three ends*, J. Differential Geom. **21** (1985), 109–127.
20. —, *The global theory of embedded minimal surfaces*, Preprint.
21. —, *One-parameter families of embedded minimal surfaces*, Preprint.
22. H. Hopf, *Lectures on differential geometry in the large*, Stanford Lecture Notes (1955); reprinted in Lecture Notes in Math., Vol. 1000, Springer-Verlag, Berlin-New York, 1984.
23. L. Jorge and W. Meeks III, *The topology of complete minimal surfaces of finite total Gaussian curvature*, Topology **22** (1983), no. 2, 203–221.
24. Dr. Krypton, *Shapes that eluded discovery*, Science Digest, April (1986).
25. N. Kuiper and W. Meeks, *Total curvature of knotted surfaces*, Invent. Math. **77** (1984), 25–69.
26. —, *The total curvature of a knotted torus*, J. Differential Geom. (to appear).
27. R. Langevin and H. Rosenberg, *On the curvature integrals and knots*, Topology **15** (1976), 405–416.
28. H. B. Lawson, Jr., *Lectures on minimal submanifolds*, Publish or Perish Press, Berkeley, 1971.
29. W. Meeks, *The topology and geometry of embedded surfaces of constant mean curvature*, J. Differential Geom. (to appear).
30. —, *The geometry and conformal structure of triply periodic minimal surfaces*, Thesis, Univ. of California at Berkeley, 1975.
31. —, *The Albanese map for a nonorientable surface*, Preprint.
32. —, *Triply periodic minimal surfaces* (in preparation).
33. W. Meeks and H. Rosenberg, *The geometry of doubly periodic minimal surfaces* (in preparation).
34. W. Meeks and S. T. Yau, *The topological uniqueness of minimal surfaces of finite topological type*, Preprint.
35. J. W. Milnor, *On the total curvature of knots*, Ann. of Math. **52** (1950), 248–257.
36. —, *On the total curvature of closed space curves*, Math. Scand. **1** (1953), 289–296.
37. R. Osserman, *A survey of minimal surfaces*, 2nd ed., Dover, New York, 1986.
38. T. Nagano and B. Smyth, *Periodic minimal surfaces*, Comment. Math. Helv. **53** (1978), 29–55.
39. B. Palmer, Ph. D. Thesis, Stanford University, 1986.
40. L. Peterson, *Three bites in a doughnut*, Science News, March (1985), 196–169.
41. A. Silveira, *Stable surfaces of constant mean curvature*, Doctoral Thesis, IMPA, Rio de Janeiro, Brasil, 1986.

- 42. R. Schoen, *Estimates for stable minimal surfaces in three-dimensional manifolds*, Seminar on Minimal Submanifolds, Princeton Univ. Press, Princeton, N.J.
- 43. —, *Uniqueness, symmetry, and embeddedness of minimal surfaces*, J. Differential Geom. **18** (1983), 791–809.
- 44. J. Spruck, *The elliptic sinh Gordan equation and the construction of toroidal soap bubbles*, Preprint.
- 45. —, personal communication.
- 46. H. Wente, *Counterexample to a conjecture of H. Hopf*, Pacific J. Math. **121** (1986), 193–243.
- 47. —, personal communication.

UNIVERSITY OF MASSACHUSETTS, AMHERST, MASSACHUSETTS 01003, USA

Almost Lie Groups

ERNST A. RUH

Lie groups play a central role in mathematics. Since their geometry is well understood, they serve as models in the study of more general manifolds. In this lecture we introduce the concept of almost Lie groups and address the question whether an almost Lie Group structure can be deformed into an exact Lie group structure.

To define almost Lie groups, let P denote a compact differentiable manifold, and let \mathfrak{g} denote a Lie algebra of the same dimension. Let $\omega: TP \rightarrow \mathfrak{g}$ denote a nondegenerate \mathfrak{g} -valued 1-form. Such a 1-form defines a parallelization of the manifold P . The vector fields $\{X_i\}$ of the parallelization are defined by $\omega(X_i) = e_i$, where $\{e_i\}$ is a basis of the Lie algebra \mathfrak{g} . The differential form ω is called a Cartan connection form, and $\Omega = d\omega + [\omega, \omega]$, where $[\ , \]$ is the Lie bracket of \mathfrak{g} , is the Cartan curvature of ω . $\Omega = 0$ is the well-known Maurer-Cartan equation which is the integrability condition for a Lie group structure with Maurer-Cartan form ω .

DEFINITION. The pair (P, ω) is called an almost Lie group if Ω is small in terms of a suitably defined norm.

The prime examples of almost Lie groups are certain subbundles P of the bundle of orthonormal frames over Riemannian manifolds. We note that almost Lie groups were introduced in Physics by Ne'eman and Regge [11, 12], under the name soft group manifolds. The purpose was to give a geometrical treatment of supergravity theories (i.e., unifying gravity with internal symmetries). In our context the question is whether or not there is a gauge transformation transforming a soft group manifold into a Lie group.

EXAMPLE. Let P denote the principal bundle of orthonormal frames over a compact Riemannian manifold M , let $\omega: TP \rightarrow \mathfrak{o}(n+1)$ be defined by $\omega = \eta + \theta$, where η is the Levi-Civita connection form, θ the canonical 1-form, and $\mathfrak{o}(n+1)$ the Lie algebra of the orthogonal group $O(n+1)$. A brief computation shows that sectional curvature K close to 1 implies $\|\Omega\|$ small and (P, ω) in an almost Lie group. A corollary of Theorem 5 in this paper establishes a gauge transformation transforming ω into a Maurer-Cartan form $\bar{\omega}$. As a consequence, M is diffeomorphic to a quotient of the sphere by a finite subgroup of $O(n+1)$.

The quality of the results depends on what is meant by $|K - 1|$ small. The best results are in low dimensions.

THEOREM 1 (HAMILTON [2]). *Every compact Riemannian manifold of dimension 3 with positive Ricci curvature is diffeomorphic to a quotient of S^3 .*

THEOREM 2 (HAMILTON [3]). *Every compact Riemannian manifold of dimension 4 with positive curvature operator is diffeomorphic to a quotient of S^4 .*

In higher dimensions the optimal theorem probably has not been proved yet. So far one has to choose between a stronger assumption and a weaker assertion.

THEOREM 3 (HUISKEN [4], MARGARIN [7], NISHIKAWA [13], RUH [15]). *Every compact Riemannian manifold of dimension n whose sectional curvature K satisfies at each point the inequality $(1 - 3(2n)^{-3/2})A \leq K \leq A$, where A is a positive function, is diffeomorphic to a quotient of S^n .*

The main ingredient in the proof of Theorems 1–3 is a heat equation for the Riemannian metric. The proof of [15] is different and does not give an explicit estimate for the pinching constant. The formulation of Theorem 3 is from [7]. [4] and [13] give similar results.

Recently a topological version of Theorem 2 was established in all dimensions.

THEOREM 4 (MICALLEF [8], MOORE [10]). *Every compact simply connected Riemannian manifold with positive curvature operator is homeomorphic to a sphere.*

The theorems of [8] and [10] are, in fact, stronger and include a pointwise quarter pinching sphere theorem and thus improve upon the celebrated sphere pinching theorems of Rauch, Berger, and Klingenberg. The proof involves the use of a complex version of the formula for the second variation of area for oriented minimal surfaces in the Riemannian manifold and the theory of holomorphic vector bundles over S^2 . Such a technique was introduced by Siu and Yau in their proof of the Frankel conjecture.

In the above example the model group $O(n + 1)$ for the almost Lie group (P, ω) is the isometry group of the standard sphere S^n . Comparison theorems for general compact symmetric spaces as models are a corollary of the following theorem on almost Lie groups of compact semisimple type.

THEOREM 5 (MIN-OO, RUH [9]). *Let \mathfrak{g} be a compact semisimple Lie algebra, $\omega: TP \rightarrow \mathfrak{g}$ a Cartan connection form on the compact manifold P with Cartan curvature Ω . There exists a positive constant A depending only on \mathfrak{g} such that $|\Omega| < A$ implies that P is diffeomorphic to a quotient Γ/G where G is the simply connected Lie group with Lie algebra \mathfrak{g} and Γ is a finite subgroup of G .*

Theorem 5 has a generalization to transversal structures of foliations. A Cartan connection ω is said to be adapted to a foliation on a manifold P if ω restricted to the leaves vanishes. An adapted connection is basic if the Cartan curvature is a basic 2-form; compare [6] for details.

THEOREM 6 (KAMBER, RUH, TONDEUR [6]). *Let \mathfrak{g} be a compact semi-simple Lie algebra, $\omega: TP \rightarrow \mathfrak{g}$ a basic Cartan connection adapted to a foliation \mathcal{F} of P with Cartan curvature Ω . There exists a positive constant A depending on \mathfrak{g} and curvature bounds on P such that $\|\kappa\|_{1,\infty} + \|\Omega\|_{1,\infty} < A$, where the mean curvature form κ of \mathcal{F} is assumed to be basic, implies that the transversal structure of \mathcal{F} is given by the simply connected Lie group G with Lie algebra \mathfrak{g} .*

The proofs of Theorems 5 and 6 (not the original proofs) can be obtained by applying the heat equation method of [2] to the Cartan connection ω . The limiting connection $\bar{\omega}$ defines the Lie group. The common feature of the above theorems is that the model structure is known a priori. This is not the case in the following generalization of the well-known Bieberbach theorem on compact Euclidean space forms.

THEOREM 7 (GROMOV [1], RUH [14]). *Let M denote a compact Riemannian manifold of dimension n , diameter d , and sectional curvature K . There exists a constant $\varepsilon = \varepsilon(n) > 0$ such that $|K|d^2 \leq \varepsilon$ implies that M is diffeomorphic to $\Gamma \backslash N$, where N is a simply connected nilpotent Lie group and Γ is an extension of a lattice $L \subset N$ by a finite group.*

The proof consists of two main steps. The first is to show that near the initial almost flat Levi-Civita connection there is a flat connection (with small torsion) and a priori bounded holonomy. This provides an almost Lie group structure $\omega: TQ \rightarrow \mathbf{R}^n$ on a finite cover Q of M . The second step is to transform the almost Lie group (Q, ω) into a Lie group. Techniques similar to the proofs of Theorems 5 and 6 can be applied. There is an additional problem: the elliptic operator associated to the heat equation has a kernel. In order to obtain a limit the Lie bracket, initially zero on \mathbf{R}^n , has to be modified in the process.

It appears to be too optimistic to expect that any almost Lie group structure can be deformed into an (integrable) Lie group structure. Theorem 7 on almost flat manifolds gives a good idea of what can be expected. Let P denote the principal bundle of orthonormal frames over an almost flat manifold M . The horizontal distribution defined by the Levi-Civita connection is almost integrable, almost a foliation. The transversal structure is given by the orthogonal group $O(n)$. An almost Lie group structure of compact semisimple type would be sufficient. Theorem 7 can be reformulated to assert that the initial almost Lie group structure can be deformed into a foliation by nilpotent Lie groups with transversal structure defined by the compact semisimple Lie group $O(n)$.

The concept of almost Lie groups can also be used to obtain results in an indirect way. Here is an example. Consider the set of Riemannian metrics with sectional curvature $|K| \leq 1$ on the 3-sphere S^3 . Let d_g denote the diameter of the Riemannian manifold (S^3, g) . Let d_{\min} denote the infimum of d_g over all metrics on S^3 with $|K| \leq 1$. Obviously, $d_{\min} \leq \pi$, the diameter of the standard S^3 . The constant $\varepsilon(3)$ of Theorem 7 on almost flat manifolds gives a lower estimate. Since $\varepsilon(3)$ is extremely small a better estimate, using the simple connectivity of S^3 , is desirable. Although the optimal answer is not yet known, the following strategy

gives a more realistic estimate. We first construct a flat metric connection ∇ on the Riemannian manifold (S^3, g) . Because S^3 is simply connected ∇ yields a parallelization, i.e., a Cartan connection $\omega: TS^3 \rightarrow \mathbb{R}^3$. The curvature bound $|K| < 1$ yields an estimate for the torsion of ∇ . If the diameter d_g were small enough, then the second step in the proof of Theorem 7 could be applied directly to $M = S^3$ leading to a contradiction. The unrealistically high estimate for the order of the holonomy group is not necessary in this case.

The construction of a flat connection ∇ on TM is based on the following observation: If, on the one hand, parallel translation along short closed loops produces orthogonal matrices with either very small or large rotations only, an averaging of sections is possible and gives a flat connection. If, on the other hand, there are orthogonal maps with intermediate size rotation, we can apply the averaging process of [5] on connections to produce a sequence of connections converging to a flat connection in TM . The intermediate size rotation produces enough cancellations in the error term to force convergence.

REFERENCES

1. M. Gromov, *Almost flat manifolds*, J. Differential Geom. **13** (1978), 231–241.
2. R. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom. **17** (1982), 255–306.
3. ———, *Four-manifolds with positive curvature operator*, Preprint, UCSD, 1985.
4. G. Huisken, *Ricci deformation of the metric on a Riemannian manifold*, J. Differential Geom. **21** (1985), 47–62.
5. H.-C. Im Hof and E. Ruh, *An equivariant pinching theorem*, Comment. Math. Helv. **50** (1975), 389–401.
6. F. Kamber, E. Ruh, and Ph. Tondeur, *Comparing Riemannian foliations with transversally symmetric foliations*, Preprint, 1986.
7. C. Margerin, *Pointwise pinched manifolds are space forms*, Proc. Sympos. Pure Math., Vol. 44, Amer. Math. Soc., Providence, R.I., 1986, pp. 307–328.
8. M. Micalef, *On the topology of positively curved manifolds*, Preprint.
9. M. Min-Oo and E. Ruh, *Comparison theorems for compact symmetric spaces*, Ann. Sci. École Norm. Sup. **12** (1979), 335–353.
10. J. Moore, *Compact Riemannian manifolds with positive curvature operators*, Bull. Amer. Math. Soc. (N.S.) **14** (1986), 279–282.
11. Y. Ne'eman and T. Regge, *Gravity and supergravity as gauge theories on a group manifold*, Phys. Lett. B **74** (1978), 54–56.
12. Y. Ne'eman and J. Thierry-Mieg, *Extended geometric supergravity on group manifolds with spontaneous fibration*, Ann. Physics **123** (1979), 243–273.
13. S. Nishikawa, *Deformation of Riemannian metrics and manifolds with bounded curvature ratios*, Proc. Sympos. Pure Math., Vol. 44, Amer. Math. Soc., Providence, R.I., 1986, pp. 343–352.
14. E. Ruh, *Almost flat manifolds*, J. Differential Geom. **17** (1982), 1–14.
15. ———, *Riemannian manifolds with bounded curvature ratios*, J. Differential Geom. **17** (1982), 643–653.

OHIO STATE UNIVERSITY, COLUMBUS, OHIO 43210, USA

Immersed Tori of Constant Mean Curvature in R^3

HENRY C. WENTE

1. Introduction. In this paper we will show how to produce a countable number of isometrically distinct immersions of a closed surface of genus one into R^3 with constant mean curvature $H \neq 0$. These examples answer a question raised by H. Hopf often posed as a conjecture.

CONJECTURE OF H. HOPF. *Let Σ be an immersed closed hypersurface with constant mean curvature $H \neq 0$ in R^n . Must Σ be the round $(n-1)$ -hypersphere?*

Two important previous results were due to A. D. Alexandroff [2] and H. Hopf [3]. Alexandroff showed that if Σ is an embedded closed hypersurface in R^n with constant mean curvature $H \neq 0$ then Σ is the round hypersphere. This improved an old result of J. H. Jellett [5] who showed the conjecture to be true if Σ is the boundary of a star-shaped region in R^3 . H. Hopf showed that an immersion of S^2 into R^3 with constant mean curvature must be a round sphere. Recently Wu-Yi Hsiang [4] constructed immersions of S^3 into R^4 with constant mean curvature different from the round hypersphere. His construction looks for surfaces invariant under the action of $O(2) \times O(2)$ on R^4 leading to an ordinary differential equation for the generating curve of the surface in the orbit space. This procedure also yields counterexamples in higher dimensions but does not work in the classical dimension ($n = 3$).

Each of our genus one examples is represented by a doubly periodic conformal immersion of R^2 into R^3 . The first examples we produce have the property that the fundamental periods for the mapping correspond to the edges of a rectangle and the preimage of the lines of curvature are straight lines parallel to the edges. There are a countable number of such examples all of which possess nice symmetry properties. We refer to these examples as the symmetric examples. In §3 we show how to construct other genus one immersions. The fundamental domain corresponding to these doubly periodic conformal immersions need not be rectangular. Furthermore, an extra rotation enters the picture leading to what we call twisted immersed tori.

2. The symmetric examples. Suppose $\bar{x}(u, v): R^2 \rightarrow R^3$ is a doubly periodic conformal immersion with constant mean curvature H . We denote points in R^2 by $(u, v) = \bar{u} = u + iv = w$ while points in R^3 are labeled $\bar{x} = (x, y, z)$. If we write the first and second fundamental forms as

$$\begin{aligned} \text{(a)} \quad ds^2 &= e^{2\omega}(du^2 + dv^2), \\ \text{(b)} \quad -d\bar{x} \cdot d\bar{\xi} &= L du^2 + 2M du dv + N dv^2 \end{aligned} \quad (2.1)$$

where $\bar{\xi}$ is the oriented unit normal to the surface then the Gauss and Codazzi equations become

$$\begin{aligned} \text{(a)} \quad \Delta\omega + Ke^{2\omega} &= 0, \\ \text{(b)} \quad \phi(w) &= [(L - N)/2] - iM \quad \text{is complex analytic.} \end{aligned} \quad (2.2)$$

Here $K = (LN - M^2)/E^2$ is the Gauss curvature with $E = e^{2\omega}$. We see immediately that $\phi(w)$ is a constant. Now $|\phi| = |k_1 - k_2|E/2$ where k_1, k_2 are the principal curvatures. If $\phi(w) \equiv 0$ then $k_1 \equiv k_2$ and the surface would be all umbilic, an impossibility for a surface genus one. So $\phi(w)$ is a nonzero constant and the immersed surface will have no umbilics. If we let $H > 0$ and let k_1 be less than k_2 , then an easy calculation gives

$$k_1 = H - |\phi|e^{-2\omega}, \quad k_2 = H + |\phi|e^{-2\omega}. \quad (2.3)$$

By a homothety of the domain we can make $|\phi| = H$ so that $k_1 = 2He^{-\omega} \sinh \omega$ and $k_2 = 2He^{-\omega} \cosh \omega$ and the Gauss equation becomes

$$\Delta\omega + 4H^2 \sinh \omega \cosh \omega = 0. \quad (2.4)$$

In general we shall select $H = \frac{1}{2}$ so that (2.4) becomes

$$\Delta\omega + \sinh \omega \cosh \omega = 0. \quad (2.5)$$

The preimage of the lines of curvature is determined by the form

$$-M du^2 + (L - N) du dv + M dv^2 = \text{Im}[\phi(w) dw^2] = 0, \quad (2.6)$$

a constant form on R^2 . The lines of curvature correspond to two families of mutually orthogonal parallel lines. The family of parallel lines corresponding to the k_1 -curvature lines will be at an angle β to the positive u -axis if $\phi(w) = -e^{-2i\beta}/2$. For our symmetric examples we set $\beta = 0$ so that the second fundamental form becomes

$$-d\bar{x} \cdot d\bar{\xi} = e^\omega [\sinh \omega du^2 + \cosh \omega dv^2]. \quad (2.7)$$

More generally the second fundamental form is

$$\begin{aligned} -d\bar{x} \cdot d\bar{\xi} &= L du^2 + 2M du dv + N dv^2, \\ L &= L(\beta) = e^\omega [(\sinh \omega) \cos^2 \beta + (\cosh \omega) \sin^2 \beta], \\ M &= M(\beta) = -\sin \beta \cos \beta, \\ N &= N(\beta) = e^\omega [(\cosh \omega) \cos^2 \beta + (\sinh \omega) \sin^2 \beta]. \end{aligned} \quad (2.8)$$

Take a doubly periodic solution to the Gauss equation (2.5) and select β for the lines of curvature, giving us the fundamental forms using (2.1) and (2.7-2.8),

both of which are doubly periodic. By a result of O. Bonnet this determines an immersion $\bar{x}(u, v)$ of R^2 into R^3 unique to within a Euclidean motion of R^3 and having the given fundamental forms. The equations to be integrated are

$$\begin{aligned} \text{(a)} \quad & \bar{x}_{uu} = \omega_u \bar{x}_u - \omega_v \bar{x}_v + L \bar{\xi}, \\ \text{(b)} \quad & \bar{x}_{uv} = \omega_v \bar{x}_u + \omega_u \bar{x}_v + M \bar{\xi}, \\ \text{(c)} \quad & \bar{x}_{vv} = -\omega_u \bar{x}_u + \omega_v \bar{x}_v + N \bar{\xi}, \\ \text{(d)} \quad & \bar{\xi}_u = (-L/E) \bar{x}_u + (-M/E) \bar{x}_v, \quad E = e^{2\omega}, \\ \text{(e)} \quad & \bar{\xi}_v = (-M/E) \bar{x}_u + (-N/E) \bar{x}_v. \end{aligned} \quad (2.9)$$

The integrated surface will be an immersed surface of constant mean curvature $H = \frac{1}{2}$ but in general will not be doubly periodic. It will have "periods."

One obtains many solutions to the Gauss equations (2.5) doubly periodic with respect to orthogonal periods from the following theorem.

THEOREM 1. *Let $\Omega(a, b) = [-a, a] \times [-b, b]$ and consider the Dirichlet problem*

$$\begin{aligned} \text{(a)} \quad & \Delta \omega + \sinh \omega \cosh \omega = 0 \quad \text{on } \Omega(a, b), \\ \text{(b)} \quad & \omega = 0 \quad \text{on } \partial \Omega(a, b). \end{aligned} \quad (2.10)$$

The system (2.10) has a unique positive solution $\omega(u, v) \equiv \omega(u, v, a, b)$ whenever the first eigenvalue for the Laplace operator with Dirichlet boundary data $\lambda_1 = (\pi^2/4)((1/a^2) + (1/b^2))$ is greater than one and no positive solution if λ_1 is less than one. The positive solutions bifurcate from the zero solution when $\lambda_1 = 1$.

By odd reflection across the grid lines $u = (2m+1)a$, $v = (2n+1)b$, the solution $\omega(u, v, a, b)$ can be extended as a doubly periodic solution to (2.5) on R^2 with half periods $\bar{p}_1 = (2a, 0)$, $\bar{p}_2 = (0, 2b)$ and whose zero set is this grid. Furthermore, the solution is symmetric about the dual grid $u = 2ma$, $v = 2nb$.

REMARKS. In [9] we analyzed (2.10) by mapping all rectangles of similar shape onto a chosen member of the family by a homothety $(u', v') = \sqrt{2\lambda}(u, v) \equiv \phi(u, v)$ and considering the system for the composed function $W(u, v) = \omega \circ \phi(u, v)$:

$$\begin{aligned} \text{(a)} \quad & \Delta W + 2\lambda \sinh W \cosh W = 0 \quad \text{on } \Omega(a_1, b_1), \\ \text{(b)} \quad & W = 0 \quad \text{on } \partial \Omega(a_1, b_1). \end{aligned} \quad (2.11)$$

In [9] we showed that there is a smooth one parameter family $W(u, v, \lambda)$ so that as $\lambda \rightarrow 0$, $W(u, v, \lambda) \rightarrow W_0(u, v, \lambda)$ where $W_0(u, v, \lambda)$ is a Green's function for $\Omega(a_1, b_1)$ with singularity in a neighborhood of the origin of the form $2 \ln(1/r)$ in polar coordinates.

Solutions to (2.10–2.11) can also be obtained by use of a "mountain-pass" argument. It has since been observed by U. Abresch [1] that the system (2.10) can be solved explicitly in closed form using elliptic functions. This fact was also used recently by R. Walter [8] to obtain integrated solutions to (2.9). In fact, the separation of variables method that enables one to solve (2.10) has been known

for some time and may be found in the book of Lamb [6], for example. Finally, the uniqueness result is due to J. Spruck [7].

Take a solution to (2.10), set $\beta = 0$, and integrate the system (2.9) to obtain our immersion $\bar{x}(u, v, a, b) \equiv \bar{x}(u, v)$. The following facts follow directly from the symmetries of the solution $\omega(u, v, a, b)$ and of the system (2.9).

I. The curve $\bar{x}(2ma, v)$ lies in a plane Π_m which is a normal plane to the surface along the curve with $\bar{x}_u(2ma, v)$ as normal vector. The mapping $\bar{x}(u, v)$ is symmetric about this plane.

$$\bar{x}(2ma - u, v) = \mathcal{R}_m \circ \bar{x}(2ma + u, v) \quad (2.12)$$

where \mathcal{R}_m is the reflection map about Π_m in R^3 .

II. The curve $\bar{x}(u, 2nb)$ lies in a normal plane Ω_n with $\bar{x}_v(u, 2nb)$ as normal vector. The mapping $\bar{x}(u, v)$ is symmetric about this plane.

$$\bar{x}(u, 2nb - v) = \mathcal{S}_n \circ \bar{x}(u, 2nb + v) \quad (2.13)$$

where \mathcal{S}_n is the reflection map about Ω_n in R^3 . The planes Π_m and Ω_n are mutually perpendicular.

III. The curve $\bar{x}(u, -b)$ lies in a plane Λ_0 which is a tangent plane to the surface along this curve. The derivative $\bar{x}_u(u, -b)$ is an even function of u about $u = a$. It follows that all the planes Π_m are parallel.

IV. The curve $\bar{x}((2m+1)a, v)$ lies on a unit sphere $S(\bar{c}_m, 1)$ with center \bar{c}_m . The centers \bar{c}_m lie on a line l which is the common intersection of the Ω_n . This line is a normal line to the planes Π_m .

The mapping $\bar{x}(u, v, a, b)$ thus exhibits a "paddle-wheel" symmetry. As a result of the discussion we can say the following. Let \bar{e} be a unit vector parallel to l . For $\bar{x}(u, v) \equiv \bar{x}(u, v, a, b)$ we have

$$\begin{aligned} \text{(a)} \quad \bar{x}(u + 4a, v) &= \bar{x}(u, v) + 2\tau\bar{e}, \\ \text{(b)} \quad \bar{x}(u, v + 4b) &= R_{2\theta} \circ \bar{x}(u, v). \end{aligned} \quad (2.14)$$

Here τ is the directed distance between two adjacent planes Π_m, Π_{m+1} while θ is the directed angle between the planes Ω_n, Ω_{n+1} . We have a map

$$\Phi(a, b) = (\tau, \theta) \quad (2.15)$$

and our surface will close up if $\tau = 0$ and $\theta/2\pi$ is rational.

THEOREM 2. *The set of all (a, b) for which $\tau(a, b) = 0$ is a curve C lying in the first quadrant of the $a - b$ plane one end of which extends to the origin. The angle function $\theta(a, b)$ is a monotonic function on C with range $0 < \theta < \pi$ so that $\theta(p) \rightarrow 0$ as $p \rightarrow (0, 0)$ along C .*

REMARKS ON PROOF. The proof of this result for small rectangles is in [9]. The argument involves a rescaling. One maps small rectangles onto a chosen rectangle of the same shape and consider solutions $W(u, v, \lambda)$ to (2.11) which determine the first fundamental form for a surface $\bar{y}(u, v, \lambda) = \bar{x}(u, v, \lambda)/\sqrt{2\lambda}$ of constant mean curvature $H = \sqrt{\lambda/2}$. As $\lambda \rightarrow 0$ the mapping $\bar{y}(u, v, \lambda)$ has a limit $\bar{y}_0(u, v)$ which is a meromorphic function. The distance function τ has a

limit as λ approaches 0. It can then be shown that the curve C extends to the origin after which one can analyze the angle function θ . The remainder of the proof has been carried out by U. Abresch [1] making use of the exact solutions.

3. Twisted immersed tori. The key to constructing more complicated examples is the following result.

THEOREM 3 [10]. *Let $\omega_0(u, v, a_0, b_0)$ be a rectangular solution to (2.10) positive on $\Omega(a_0, b_0)$ and having the symmetry properties listed in Theorem 1. Suppose that it satisfies the following stability condition. The only solution to*

$$\begin{aligned} \text{(a)} \quad & \Delta h + (\cosh 2\omega)h = 0 \quad \text{on } \Omega(a_0, b_0), \\ \text{(b)} \quad & h = 0 \quad \text{on } \partial\Omega(a_0, b_0) \end{aligned} \quad (3.1)$$

is $h \equiv 0$. Then there is a neighborhood $\mathcal{U} \subset R^3$ of $(a_0, b_0, 0)$ and a smooth function $\omega(u, v, a, b, c): R^2 \times \mathcal{U} \rightarrow R$ such that

- (i) *for each $(a, b, c) \in \mathcal{U}$, $\omega(u, v, a, b, c)$ is a solution to (2.5) on R^2 ;*
- (ii) *if $\bar{p}_1 = (2a, 0)$ and $\bar{p}_2 = (2c, 2b)$, then $\omega(\bar{u} + \bar{p}_i, \bar{a}) = -\omega(\bar{u}, \bar{a})$ where $\bar{u} = (u, v)$ and $\bar{a} = (a, b, c)$. We say that $\omega(\bar{u}, \bar{a})$ is half period odd;*
- (iii) $\omega(\bar{u}, \bar{a}) = \omega(-\bar{u}, \bar{a})$;
- (iv) $\omega(u, v, a_0, b_0, 0) = \omega_0(u, v, a_0, b_0)$.

We have a three-parameter family $\omega(u, v, a, b, c)$ of solutions to (2.5) with periods $2\bar{p}_1 = (4a, 0)$ and $2\bar{p}_2 = (4c, 4b)$. We let β determine the second fundamental form and integrate (2.9) to obtain a map $\bar{x}(u, v, a, b, c, \beta)$, a constant mean curvature immersion with $H = \frac{1}{2}$ of R^2 into R^3 . Because of the periodicity of the first and second fundamental forms we infer that

$$\bar{x}(\bar{u} + 2\bar{p}_i, a, b, c, \beta) = \mathcal{E}_i \circ \bar{x}(\bar{u}, a, b, c, \beta), \quad i = 1, 2, \quad (3.2)$$

where $\bar{p}_1 = (2a, 0)$, $\bar{p}_2 = (2c, 2b)$ and $\mathcal{E}_1, \mathcal{E}_2$ are commuting Euclidean motions in R^3 . We note the following result.

THEOREM 4 [10]. *Let $\mathcal{E}_1, \mathcal{E}_2$ be commuting Euclidean motions in R^3 where*

$$\mathcal{E}_i \circ \bar{x} = A_i \bar{x} + \bar{a}_i, \quad A_i \in \text{SO}(3), \quad \bar{a}_i \in R^3. \quad (3.3)$$

Suppose that $A_1 \neq I$ and A_2 is not a 180° -rotation matrix. There is a unique common one-dimensional eigenspace with eigenvalue one for A_1 and A_2 . Let \bar{e} be a unit eigenvector for this eigenspace. There is a unique vector \bar{p} orthogonal to \bar{e} so that

$$\mathcal{E}_i \bar{x} = A_i(\bar{x} - \bar{p}) + \bar{p} + \tau_i \bar{e}. \quad (3.4)$$

Let l be the line through \bar{p} with direction vector \bar{e} . We may rewrite (3.4) in the form

$$\mathcal{E}_i \bar{x} = \mathcal{R}_{\theta_i} \circ \bar{x} + \tau_i \bar{e} \quad (3.5)$$

where \mathcal{R}_{θ_i} is a rotation about the line l through an angle θ_i (determined mod 2π).

Theorem 4 applies to our situation. Take a solution $\omega(u, v, a, b, c)$ as determined by Theorem 3, choose the angle β , and construct our constant mean

curvature immersion satisfying (3.2). We obtain a smooth map $\Phi: \mathcal{U} \times R \rightarrow R^4$ where \mathcal{U} is a neighborhood of $(a_0, b_0, 0)$ in R^3 ,

$$\Phi(a, b, c, \beta) = (\tau_1, \theta_1, \tau_2, \theta_2). \quad (3.6)$$

THEOREM 5 [10]. *Let $(a_0, b_0, 0, 0)$ be chosen so that the corresponding symmetric immersion gives a closed immersed torus. That is, $\tau_1 = \tau_2 = 0$, $\theta_1 = 0$, and $\theta_2 = 2\pi r_0$ where r_0 is rational. For almost all (perhaps all) such immersions the map Φ given by (3.6) is locally invertible. If we set*

$$(D\Phi)^T = \begin{pmatrix} \Phi_a \\ \Phi_b \\ \Phi_c \\ \Phi_\beta \end{pmatrix} = \begin{pmatrix} A & 0 & 0 & B \\ C & 0 & 0 & D \\ 0 & E & F & 0 \\ 0 & 0 & G & 0 \end{pmatrix} \quad (3.7)$$

where Φ_a is a row vector, then $AD - BC \neq 0$, $E \neq 0$, $G \neq 0$ at $(a_0, b_0, 0, 0)$.

We observe that if $c = \beta = 0$ then we are in the class of symmetric immersions and $\Phi(a, b, 0, 0) = (\tau_1, 0, 0, \theta_2)$. This explains the 0-entries in the first two rows of (3.7). The 0-entries in the first and fourth columns of rows three and four can also be explained by symmetry considerations. The bulk of the work in proving Theorem 5 is in showing $E \neq 0$, $G \neq 0$, and that the entry in row four, column two, is zero.

Our starting point is a 4-tuple $(a_0, b_0, 0, 0)$ with $\Phi(a_0, b_0, 0, 0) = (0, 0, 0, 2\pi r_0)$ where r_0 is rational. Theorem 5 says that Φ has a local inverse. Thus for all $(0, 2\pi s, 0, 2\pi r)$ with r, s rational, r near r_0 , and s near 0, there is an immersed torus of constant mean curvature where $\Phi(a, b, c, \beta) = (0, 2\pi s, 0, 2\pi r)$. As one increments through a period $2\bar{p}_1 = (4a, 0)$ or $2\bar{p}_2 = (4c, 4b)$ of the function $\omega(u, v, a, b, c)$, there will be rotations $\theta_1 = 2\pi s$ and $\theta_2 = 2\pi r$ in the immersion $\bar{x}(u, v, a, b, c, \beta)$. We have constructed an enlarged class of twisted immersed tori.

Acknowledgments. This work was supported, in part, by a grant from the National Science Foundation and by visits, while on sabbatical, to the University of California, San Diego, and to the University of Bonn, Germany, with support from SFB 72.

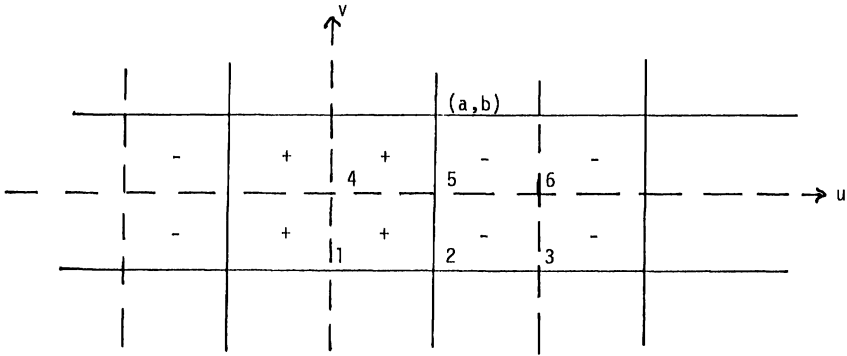
Illustrations.

FIGURE 1. The zero set $u = (2m + 1)a$, $v = (2n + 1)b$ for the symmetric solution $\omega(u, v) \equiv \omega(u, v, a, b)$ to (2.5). The function is symmetric about the dual grid $u = 2ma$, $v = 2nb$.

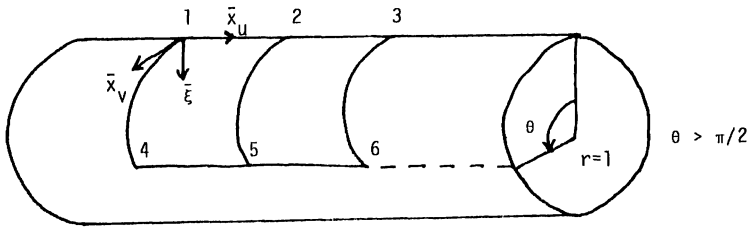


FIGURE 2. The surface $\bar{x}(u, v, a, b)$ when $\omega(u, v, a, b) \equiv 0$. (A pure cylinder)

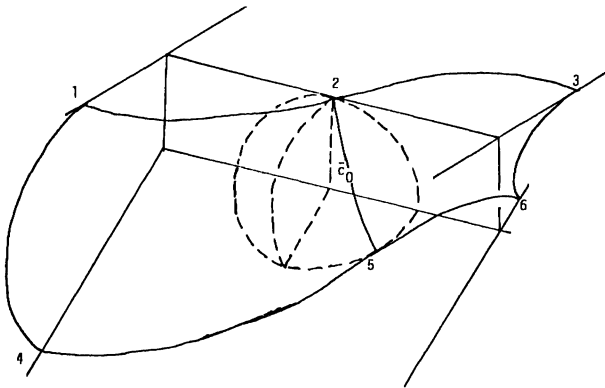


FIGURE 3. A section of the image surface $\bar{x}(u, v, a, b)$ when $\omega(u, v, a, b)$ is nonzero. The planes Π_m are still separated.

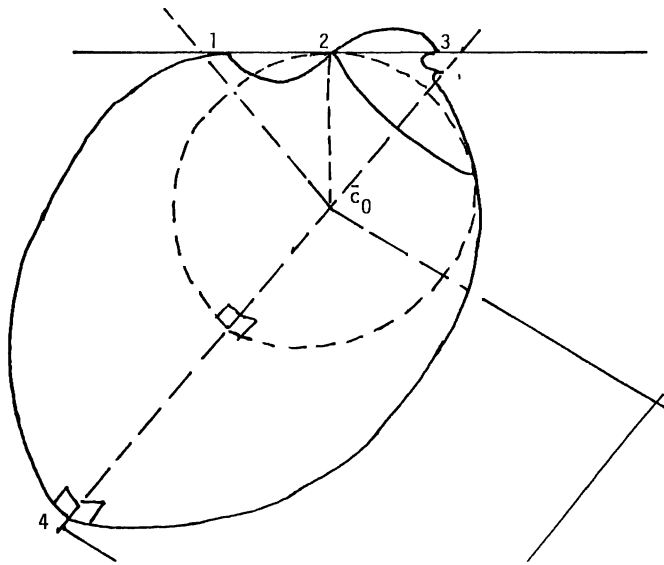


FIGURE 4. A section of the surface when the planes Π_m have coalesced. Reflect this figure about the plane of the paper to obtain a section of surface shaped like a "clam shell." Now rotate this figure 180° about a vertical line through \bar{c}_0 . The figure now resembles a clam with its mouth open. A closed surface is obtained by a succession of reflections.

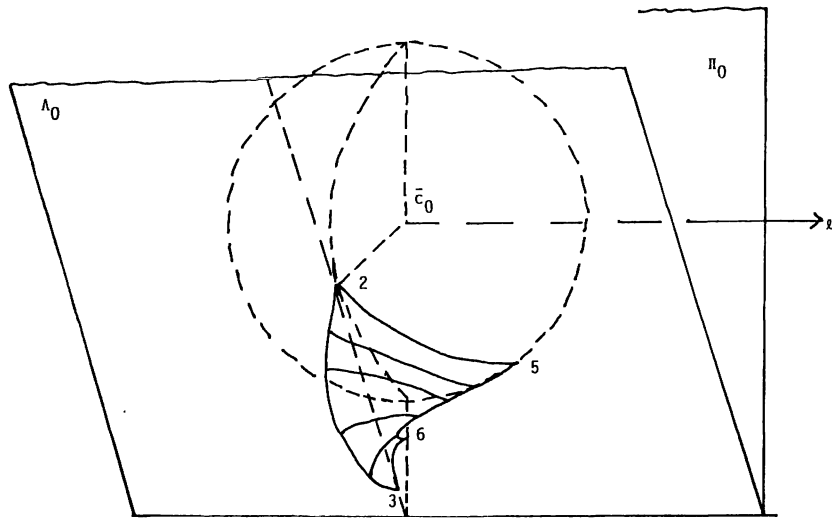


FIGURE 5. The image of a rectangular zone of negative Gauss curvature resembles an Enneper's surface. A closed surface may be thought of as sphere-like objects held together by pieces of Enneper's surface.

REFERENCES

1. U. Abresch, *Constant mean curvature tori in terms of elliptic functions*, Preprint, Max-Planck Institut fur Mathematik, Bonn, 1985.
2. A. D. Alekandrov, *Uniqueness theorems for surfaces in the large*, Vestnik, Leningrad. Univ. **13** (1958), no. 19, 5-8; English transl. in Amer. Math. Soc. Transl. (2) **21** (1962), 412-416.
3. H. Hopf, *Differential geometry in the large*, Lecture Notes in Math., vol. 1000, Springer, 1983.
4. Wu-Yi Hsiang, *Generalized rotational hypersurfaces of constant mean curvature in the Euclidean spaces. I*, J. Differential Geom. **17** (1982), 337-356.
5. J. H. Jellett, *Sur la surfaces dont la courbure moyenne est constante*, J. Math. Pures Appl. **18** (1853), 163-167.
6. G. L. Lamb, Jr., *Elements of soliton theory*, Wiley, New York, 1980.
7. J. Spruck, *The elliptic sinh Gordon equation and the construction of toroidal soap bubbles*, Preprint, 1986.
8. R. Walter, *Explicit examples of the H-problem of H. Hopf*, Preprint, 1986.
9. H. C. Wente, *Counterexample to a conjecture of H. Hopf*, Pacific J. Math. **121** (1986), no. 1, 193-243.
10. —, *Twisted tori of constant mean curvature in R^3* , Tromba, Seminar on New Results in Nonlinear Partial Differential Equations, Vieweg, 1987.

UNIVERSITY OF TOLEDO, TOLEDO, OHIO 43606, USA

Segal's Burnside Ring Conjecture and Related Problems in Topology

GUNNAR CARLSSON

In 1960, M. F. Atiyah proved the following theorem.

THEOREM (ATIYAH, [2]). *Let G be a finite group, and let U denote the infinite unitary group. Then $[BG, BU \times \mathbf{Z}]$ is isomorphic to the completion of the complex representation ring $R[G]$ at its augmentation ideal. (Here $[X, Y]$ denotes the homotopy classes of maps from X to Y , and if G is any topological group, BG denotes its classifying space.)*

This theorem is motivated by the observation that any complex representation of G , i.e., homomorphism $\rho: G \rightarrow U(n)$, induces a map

$$\theta(\rho): BG \xrightarrow{B\rho} BU(n) \rightarrow BU \times n \rightarrow BU \times \mathbf{Z},$$

and that if two representations are isomorphic, they induce homotopic maps. It is natural to ask if all maps $f: BG \rightarrow BU$ are in the image of θ , i.e., are homotopic to maps induced by representations. This turns out to be false, but Atiyah's theorem very precisely describes the failure of this naive guess. One first observes that $BU \times \mathbf{Z}$ admits a homotopy commutative product, arising from Whitney sum of vector bundles, so that $[BG, BU \times \mathbf{Z}]$ becomes an abelian group. Furthermore, if ρ and ρ' are representations, one has the formula $\theta(\rho \oplus \rho') = \theta(\rho) + \theta(\rho')$, so that θ extends over the group completion of the abelian monoid $\text{Rep}[G]$ of isomorphism classes of representations of G , which is $R[G]$, giving a homomorphism $\bar{\theta}: R[G] \rightarrow [BG, BU]$. Atiyah then shows that $\bar{\theta}$ actually extends over the completion $\hat{R}[G]$ of $R[G]$ at its augmentation ideal, and that this extension is an isomorphism. Thus, one can informally say that the homotopy classes of maps induced by representations are dense among all maps from BG to BU .

It is very surprising to find such a neat, closed-form expression for the problem of classifying up to homotopy maps from BG to BU ; in general, the set $[X, Y]$ is

Supported in part by National Science Foundation Grant DMS-86-02430. The author is an Alfred P. Sloan Foundation Fellow.

notoriously difficult to describe. This led topologists to hope that similar closed-form descriptions might be available in certain other cases, where one replaces the space BU by a different target space. Graeme Segal, in particular, considered the space $Q(S^0)$. Recall that this space is defined as the direct limit of the spaces of based loops $\Omega^n S^n$, where $\Omega^n X$ denotes the based maps of the space S^n into X . The direct limit system is then defined by suspension of maps, using the fact that $\Sigma S^n = S^{n+1}$. Segal asked if there was a reasonable description of the set (in fact, abelian group) $[BG, Q(S^0)]$. To see what the statement analogous to Atiyah's should be, we must describe the so-called Dyer-Lashof maps. Consider the space $C_n(\mathbf{R}^k) = \{(v_1, \dots, v_n) \mid v_i \in \mathbf{R}^k, d(v_i, v_j) \geq 1 \text{ for } i \neq j\}$; d denotes Euclidean distance. This is the space of all ordered distinct n -tuples in \mathbf{R}^k , so that the distance between any two distinct points is greater than or equal to one. We may view S^k as the one-point compactification $\mathbf{R}^k \cup \{\infty\}$, or as the one-point compactification $B \cup \{\infty\}$, where B is any open ball in \mathbf{R}^k . Given any $p \in \mathbf{R}^k$, define the associated collapse map $\phi_p: \mathbf{R}^k \cup \{\infty\} \rightarrow B_{1/2}(p) \cup \{\infty\}$ by

$$\begin{aligned} \phi_p(x) &= \infty & \text{if } p \notin B_{1/2}(p), \\ \phi_p(x) &= x & \text{if } p \in B_{1/2}(p). \end{aligned}$$

Given an element $(p_1, \dots, p_n) \in C_n(\mathbf{R}^k)$, we define a map

$$\psi_{(p_1, \dots, p_n)}: S^k \simeq \mathbf{R}^k \cup \{\infty\} \rightarrow \bigvee_{i=1}^n (B_{1/2}(p_i) \cup \{\infty\}) \cong \bigvee_{i=1}^n S^k$$

by

$$\begin{aligned} \psi_{(p_1, \dots, p_n)}(x) &= \phi_{p_i}(x) & \text{if } x \in B_{1/2}(p_i), \\ \psi_{(p_1, \dots, p_n)}(x) &= \infty & \text{if } x \notin \bigcup_i B_{1/2}(p_i). \end{aligned}$$

We now define the Dyer-Lashof map $D_{n,k}: C_n(\mathbf{R}^k) \rightarrow \Omega^k S^k$ by $D_{n,k}(p_1, \dots, p_n) = F \circ \psi_{(p_1, \dots, p_n)}$, where $F: \bigvee_{i=1}^n S^k \rightarrow S^k$ is the evident folding map. We obtain the commutative diagram

$$\begin{array}{ccc} C_n(\mathbf{R}^k) & \xrightarrow{D_{n,k}} & \Omega^k S^k \\ \downarrow & & \downarrow \\ C_n(\mathbf{R}^{k+1}) & \xrightarrow{D_{n,k+1}} & \Omega^{k+1} S^{k+1} \end{array}$$

and hence we obtain a map $D_n: \bigcup_k C_n(\mathbf{R}^k) \rightarrow Q(S^0)$. One can show that $\bigcup_k C_n(\mathbf{R}^k)$ is a contractible space. Note also that Σ_n acts freely on $\bigcup_k C_n(\mathbf{R}^k)$, and that the map D factors through $\bigcup_k C_n(\mathbf{R}^k)/\Sigma_n$. Since $\bigcup_k C_n(\mathbf{R}^k)$ is contractible and Σ_n acts freely, this orbit space is a classifying space for Σ_n , i.e., $B\Sigma_n$. So D_n defines a map $D_n: B\Sigma_n \rightarrow Q(S^0)$. It turns out that D_n has its image in the component of $Q(S^0)$ consisting of maps of degree n ; $Q(S^0)$ is an H -space, so we may add any fixed map of degree $-n$ and obtain a map $\tilde{D}_n: B\Sigma_n \rightarrow Q(S^0)_0$, the component of maps of degree 0. These maps are compatible over n and produce a map $\tilde{D}_\infty: B\Sigma_\infty \rightarrow Q(S^0)_0$. This map is proved by Barratt, Priddy, and Quillen [9] to induce an isomorphism on integral homology; in fact, one can show, using simplicial techniques, that $Q(S^0)$ may be

constructed from the space $B\Sigma_\infty$ in a very explicit manner. This suggested to Segal that representations of a finite group into Σ_n should play the role that complex representations do for BU . The next step is to determine what is the "representation ring" for representations into the symmetric group. This turns out to be a well-known construction, due to Dress, called the *Burnside ring* of G . It is usually written $A(G)$. It is defined as follows. Recall that a G -set is simply a set together with a G -action on the set. The notion of isomorphism of G -sets is the evident one. Let $\mathcal{S}(G)$ denote the set of isomorphism classes of finite G -sets; then $\mathcal{S}(G)$ is a commutative monoid under the sum operation $[X_1] + [X_2] = [X_1 \amalg X_2]$, where \amalg denotes disjoint union of G -sets. $A(G)$ as an abelian group is the group completion of this monoid. The product operation $[X_1] \cdot [X_2] = [X_1, X_2]$ yields a ring structure on $A(G)$. $A(G)$ is always a free abelian group whose rank is equal to the number of conjugacy classes of subgroups of G ; this comes from the fact that every finite G -set X has a unique decomposition $X = \bigcup_i G/K_i$, where G/K_i denotes the G -set of right K_i -cosets in G . $A(G)$ is also equipped with an augmentation $A(G) \rightarrow A(\{e\}) \cong \mathbb{Z}$, viewing a G -set as a set with no group action. The kernel of this map is called $I(G)$, the augmentation ideal of $A(G)$. Segal observed that the Dyer-Lashof construction yields a homomorphism from $A(G)$ to $[BG, Q(S^0)]$, and that this homomorphism extends to a homomorphism $\theta: \hat{A}(G) \rightarrow [BG, Q(S^0)]$, where $\hat{A}(G)$ denotes the completion of $A(G)$ at $I(G)$. He made the following conjecture.

CONJECTURE (SEGAL). *θ is an isomorphism.*

This conjecture has recently been proved. The proof can be broken up into four steps.

(I) *Prove that the conjecture holds for cyclic groups of prime order.* This was carried out first for $p = 2$ by W. H. Lin [6], using difficult Adams spectral sequence techniques. J. H. C. Gunawardena [5] adapted these techniques to the case of an odd prime p .

(II) *Prove that the conjecture holds for $G = (\mathbb{Z}/p\mathbb{Z})^k$.* This was proved by Adams, Gunawardena, and Miller [1], again using the Adams spectral sequence. The technique here is a vast generalization and refinement of Lin's techniques.

(III) *Prove that the conjecture holds for a general p -group if it holds whenever $G \cong (\mathbb{Z}/p\mathbb{Z})^k$.* This was proved by the author [4]. The key ingredients are a reformulation of the conjecture in terms of "equivariant stable homotopy theory," the construction of fixed point free representations with nilpotent Euler class on a nonelementary abelian p -group G , using the Quillen-Venkov theorem [11], and the contractibility of the classifying space of the ordered set of all proper, non-trivial subgroups of a nonelementary abelian p -group G . This is also a theorem of Quillen [10]. The equivariant reformulation actually permits the analysis of the groups $[\Sigma^k BG, Q(S^0)]$ for all k as well; see [4] for details.

(IV) *Prove that the conjecture holds for a general finite group if it holds for all p -groups.* This was proved by May and McClure [7], by showing that the functor $\hat{A}(G)$ satisfies sufficiently strong induction conditions to give the result.

An interesting feature of the proof is that it is very distinct in character from the proof of Atiyah's theorem. Atiyah relied heavily on the computability of the groups $\pi_n(BU)$, and used an induction based on the Atiyah-Hirzebruch spectral sequence. On the other hand, the groups $\pi_n(Q(S^0))$ are the "stable homotopy groups of spheres," groups whose description remains an open problem; therefore, they cannot be used to ground an induction. The proof of Segal's conjecture uses a smaller amount of computational input, and can be modified to give a short proof of Atiyah's theorem.

As it stands, this theorem is a very pleasing analogue to Atiyah's theorem. It turns out, though, that it can be applied to study a family of problems of central importance in algebraic topology. Let G be a group, and let EG be a contractible space on which G acts freely. One can show that such a space always exists; $EG/G \cong BG$. Let X be a G -space. Then one can view the function space $F(EG, X)$ as a G -space by setting $g \cdot f = gfg^{-1}$. The fixed point set $F(EG, X)^G$ is referred to as the homotopy fixed point set. On the other hand, let p denote the one-point space with trivial G -action. Then $F(p, X)$ can also be viewed as a G -space; of course, it is canonically G -homeomorphic to X . The map $EG \rightarrow p$ determines a G -map $X \cong F(p, X) \rightarrow F(EG, X)$, and hence a map $X^G \rightarrow F(EG, X)^G$. The "homotopy limit problem" asks for an analysis of this map, hopefully proving that it is an equivalence in some sense. R. Thomason [14] introduced this class of problems. Theorems showing an equivalence between fixed point sets and homotopy fixed point sets are of a great use for the following reasons. The first is that, in some contexts, the fixed point set is easily described, and one then obtains a description of an otherwise intractable homotopy fixed point set. The second is that by filtering EG by its skeleta, one obtains a spectral sequence with $E_2^{p,q} \cong H^p(G; \pi_{-q}(X))$ converging to $\pi_*(F(EG, X)^G)$. This gives information in situations where the fixed point set is intractable. I will illustrate this with two examples.

The first is Sullivan's conjecture, as proved by Haynes Miller [8]. It asks if the based mapping space $F(BG, X)$ is contractible, where G is a finite group, and X is a finite complex. Equivalently, one could ask that the unbased mapping space should be weakly equivalent to X . Now, if we view X as a G -complex with trivial action, $F(BG, X) \cong F(EG, X)^G$. Thus, Sullivan's conjecture could be stated as asking if the fixed point set and homotopy fixed point set of a finite complex with trivial action are equivalent. One could, of course, ask the question with nontrivial action; this is the generalized Sullivan conjecture.

An example of the second type of situation where homotopy limit problems are useful is the conjectured existence of a descent spectral sequence for algebraic K -theory. Since Quillen invented his algebraic K -groups of rings, their calculation in many situations of interest has remained an unsolved problem. In particular, the case of fields is a particularly central case, and the only complete calculation is that of Suslin [13], who computes the algebraic K -theory with finite coefficients of an algebraically closed field. Nonalgebraically closed fields, such as \mathbf{Q} or number fields, remain a mystery. A possible approach is the following. The

algebraic K -theory of a field L can be described as the homotopy groups of a certain space $BGL^+(L)$. If L is a Galois extension of a subfield L_0 , with Galois group G , then G acts on $BGL^+(L)$, with fixed point set $BGL^+(L_0)$. If one could prove that the fixed point set was equal to the homotopy fixed point set in this situation, one would have a spectral sequence with $E_2^{p,q} = H^p(G; K_{-q}(L))$, converging to $K_*(L_0)$, which would be a big step toward the solution of the problem.

It turns out that the Segal conjecture can be viewed as a homotopy limit problem. The space X is now $Q^G(S^0)$, a G -space constructed by Segal whose underlying space is $Q(S^0)$, but which is equipped with a group action coming from all the linear actions on spheres. See [12] for details of this construction. It further turns out that the Segal conjecture can be used to attack the generalized Sullivan conjecture and the descent problem for algebraic K -theory, in the case of number fields, p -adic local fields, and function fields. The method used is to approximate the G -space X in question by a tower of G -fibrations

$$\begin{array}{ccc} & & \downarrow \\ F_2 & \rightarrow & X_2 \\ & & \downarrow \\ F_1 & \rightarrow & X_1 \\ & & \downarrow \\ & & X_0 \\ & & \downarrow \\ & & * \end{array}$$

so that X_0 and F_i are G -spaces for which the homotopy limit problem has an affirmative solution by the Segal conjecture. This tower of fibrations arises from the so-called “cosimplicial space of the triple Q^G ”; see [3]. In the case of the generalized Sullivan conjecture, one obtains a complete description of the homotopy type of $F(EG; X)^G$ at the prime p , if p is a p -group. In the case of the descent spectral sequence for algebraic K -theory, one obtains enough information about finite descent to begin a serious attack on the case of descent from $\overline{\mathbb{Q}}$, the field of all algebraic numbers, to an algebraic number field. This descent is more difficult, since the Galois group is profinite rather than finite.

Finally, I should state that in the case of the Sullivan conjecture, H. R. Miller has obtained, independently and simultaneously, a proof of essentially the same theorem; J. Lannes has since provided another proof.

REFERENCES

1. J. F. Adams, J. H. C. Gunawardena, and H. R. Miller, *The Segal conjecture for elementary abelian p -groups*, *Topology* **24** (1985), 435–460.
2. M. F. Atiyah, *Characters and cohomology of finite groups*, *Inst. Hautes Études Sci. Publ. Math.* **9** (1961), 23–64.
3. A. K. Bousfield and D. M. Kan, *Homotopy limits, completions, and localizations*, *Lecture Notes in Math.*, vol. 304, Springer-Verlag, 1972.

4. G. Carlsson, *Equivariant stable homotopy theory and Segal's Burnside ring conjecture*, Ann. of Math. (2) **120** (1984), 189–224.
5. J. H. C. Gunawardena, *Segal's conjecture for cyclic groups of (odd) prime order*, J. T. Knight Prize Essay, Cambridge University, 1980.
6. W. H. Lin, *On conjectures of Mahowald, Segal, and Sullivan*, Math. Proc. Cambridge Philos. Soc. **87** (1980), 449–458.
7. J. P. May and J. E. McClure, *A reduction of the Segal conjecture* (London, Ontario, 1981), CMS Conf. Proc., vol. 2, part 2, Amer. Math. Soc., Providence, R.I., 1982, pp. 209–222.
8. H. R. Miller, *The Sullivan conjecture on maps from classifying spaces*, Ann. of Math. (2) **120** (1984), 39–87.
9. S. B. Priddy, *On $\Omega^\infty S^\infty$ and the infinite symmetric groups*, Proc. Sympos. Pure Math., vol. 22, Amer. Math. Soc., Providence, R.I., 1971, pp. 217–220.
10. D. G. Quillen, *Homotopy properties of the poset of non-trivial p -subgroups of a group*, Adv. in Math. **28** (1978), 101–128.
11. D. G. Quillen and B. Venkov, *Cohomology of finite groups and elementary abelian subgroups*, Topology **11** (1972), 552–568.
12. G. B. Segal, *Equivariant stable homotopy theory*, Proc. Internat. Congr. Math. (Nice, 1970), vol. 2, Gauthier-Villars, Paris, 1971, pp. 59–63.
13. A. A. Suslin, *On the K -theory of algebraically closed fields*, Invent. Math. **73** (1983), 241–245.
14. R. Thomason, *The homotopy limit problem*, Proceeding Northwestern Homotopy Theory Conference (Evanston, Ill., 1982), Contemp. Math., vol. 19, Amer. Math. Soc., Providence, R.I., 1983, pp. 407–420.

PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08540, USA

The Sullivan Conjecture and Homotopical Representation Theory

HAYNES MILLER

In his well-known M.I.T. notes [39, 5.115–5.131] Dennis Sullivan made a conjecture about the mod 2 homology type of the fixed point set of an involution. As explained below, this conjecture may be regarded as a geometric version of the classical P. A. Smith theory. In a special case, it led Sullivan to a further conjecture about maps from real projective spaces.

Forms of these conjectures have recently been proven, and they have found diverse applications in homotopy theory. In time they may find application in geometric topology as well; cf. [37, 5]. The object of this talk is to describe these conjectures in their original context (§1), then indicate some of the methods of proof and related consequences of these methods (§2), and finally (§3) to indicate some applications to the study of maps between classifying spaces.

The Sullivan fixed point conjecture is a case of a much more general issue, the “homotopy limit problem.” As a general discussion of this problem may be found in G. Carlsson’s talk, we will not dwell on this aspect of the matter here.

1. The Sullivan conjectures. One may be stated as

THEOREM [31]. *Let X be a finite-dimensional CW complex and let $f: \mathbf{R}P^n \rightarrow X$ be a continuous map. If f can be extended over $\mathbf{R}P^{n+k}$ for arbitrary k , then it is homotopic to a constant map.*

This theorem may be conveniently packaged by forming $\mathbf{R}P^\infty = \bigcup \mathbf{R}P^n$. Then it is contained in

THEOREM [31]. *With X as above, $\text{map}_*(\mathbf{R}P^\infty, X)$, the space of pointed maps with its compact-open topology, is weakly contractible: all its homotopy groups are trivial.*

As a result, if K is any cell-complex, there are no essential maps from $\mathbf{R}P^\infty$ to $\text{map}_*(K, X)$, despite the fact that the mapping space is usually very large (a loop-space, for instance).

Since $\mathbf{R}P^\infty$ has only one nontrivial homotopy group ($\pi_1(\mathbf{R}P^\infty) = \mathbf{Z}_2$), this theorem is a case of

THEOREM 1 (A. ZABRODSKY [41]; SEE ALSO [31]). *Let W be a CW complex with only finitely many nonzero homotopy groups, each finite (or merely locally finite). Let X be any finite-dimensional CW complex. Then $\text{map}_*(W, X)$ is weakly contractible.*

In particular, we may have $W = B\pi$, the classifying space of a (locally) finite group π ; in fact this case, proved in [31], starts an induction for Zabrodsky. In [41] he also allows the source to have nontrivial rational homotopy, under some further side conditions. He shows that all maps are then phantom maps, and that $\pi_*\text{map}_*(W, X)$ is computable in terms of $H_*(W; \mathbf{Q})$ and $\pi_*(X) \otimes \mathbf{Q}$. In [18], E. M. Friedlander and G. Mislin apply Theorem 1 and étale methods to the study of $\text{map}_*(BG, X)$ for G a compact Lie group and X a finite complex. Another application of this work is the affirmation by C. McGibbon and J. A. Neisendorfer [29] of a conjecture of J.-P. Serre: If X is a simply connected finite complex with $\overline{H}_*(X; \mathbf{Z}_p) \neq 0$, then $\pi_i(X)$ contains p -torsion for infinitely many i . (This theorem is proven anew, without the use of Theorem 1 or the like, by J. Lannes and L. Schwartz in [23].) See [45] for further applications.

To put these results in context one should consider a more general situation.

Let π be a group acting on a space X . The *homotopy fixed point set* is the space

$$X^{h\pi} = \text{map}^\pi(E\pi, X)$$

of equivariant maps to X from a free contractible π -CW complex $E\pi$. The actual fixed point set X^π embeds in $X^{h\pi}$ as the subspace of constant maps. The functor $X \mapsto X^{h\pi}$ has the following homotopy property, not enjoyed by $X \mapsto X^\pi$.

(H) Let $f: X \rightarrow Y$ be an equivariant map. If f is a homotopy equivalence, then so is the induced map $f^{h\pi}: X^{h\pi} \rightarrow Y^{h\pi}$.

How close is $X^\pi \rightarrow X^{h\pi}$ to being an equivalence? It is elementary to show:

EXAMPLE 1. If F is a free π -CW complex (e.g., π itself, with π acting by translation) and $X = \text{map}(F, Y)$, with π acting by conjugation and Y arbitrary, then $X^\pi \rightarrow X^{h\pi}$ is a homotopy equivalence.

EXAMPLE 2. Suppose that π acts freely on X and that X is of finite category (e.g., is finite-dimensional). If π has any nontrivial torsion, then $X^{h\pi} = \emptyset (= X^\pi)$.

Theorem 1 gives us

EXAMPLE 3. Suppose that π acts trivially on X . Then $X^{h\pi} = \text{map}(B\pi, X)$, and $X^\pi = X$; so if X is a finite-dimensional CW complex and π is locally finite, then $X^\pi \rightarrow X^{h\pi}$ is a weak homotopy equivalence.

In [21] L. Jones established the following converse to Smith theory. If F is a finite complex with $\overline{H}_*(F; \mathbf{Z}_p) = 0$, then there is a contractible finite \mathbf{Z}_p -CW complex X with $X^{\mathbf{Z}_p} = F$. Since by (H) $X \rightarrow *$ induces a homotopy equivalence $X^{h\mathbf{Z}_p} \rightarrow *$, the best we can expect of $X^\pi \rightarrow X^{h\pi}$ in general is good

homological behavior; and indeed, in [39] Sullivan proposed essentially

CONJECTURE (SC). If X is a finite-dimensional \mathbf{Z}_2 -CW complex, then $X^{\mathbf{Z}_2} \rightarrow X^{h\mathbf{Z}_2}$ induces an isomorphism in mod 2 homology.

This conjecture arose in the context of Sullivan's use of the étale homotopy theory of M. Artin and B. Mazur [4] to prove the Adams conjecture. We give a somewhat updated account of his chain of reasoning, because it serves to motivate the form of Theorem 2.

Let V be an algebraic variety (to be precise, a pointed connected scheme of finite type) over \mathbf{C} . Étale homotopy theory (as presented in [17]; see Corollary 8.5) provides a model $V_{\text{ét}}$ for the analytic space $V(\mathbf{C})$ of complex points, constructed purely from the structure of V as algebraic variety, without using the topology of \mathbf{C} . By specializing to the 2-adic part, one has a map $V(\mathbf{C}) \rightarrow V_{\text{ét},2}$ which is up to homotopy the same as the \mathbf{Z}_2 -completion $V(\mathbf{C}) \rightarrow \mathbf{Z}_{2\infty}V(\mathbf{C})$ of Bousfield and Kan [6]. We pause to record some properties of this general construction $\mathbf{Z}_{p\infty}$, p a prime.

(BK1) $X \rightarrow \mathbf{Z}_{p\infty}X$ is a functor from spaces to spaces (simplicial sets to simplicial sets, actually), and comes with a natural transformation $\eta: X \rightarrow \mathbf{Z}_{p\infty}X$.

(BK2) If X is simply connected then so is $\mathbf{Z}_{p\infty}X$, and η is a Bousfield $H_*(-; \mathbf{Z}_p)$ -localization [7]. This means (1) $H_*(\eta; \mathbf{Z}_p)$ is an isomorphism, and (2) any mod p homology equivalence $A \rightarrow B$ induces a bijection $[A, \mathbf{Z}_{p\infty}X] \leftarrow [B, \mathbf{Z}_{p\infty}X]$.

(BK3) Let $f: X \rightarrow Y$. Then $H_*(f; \mathbf{Z}_p)$ is an isomorphism if and only if $\mathbf{Z}_{p\infty}f$ is a homotopy equivalence.

Now suppose V is defined over \mathbf{R} . Then $V(\mathbf{C})$ carries a continuous involution with fixed point subspace $V(\mathbf{R})$. Sullivan now wondered whether $V(\mathbf{R})$ —or rather $\mathbf{Z}_{2\infty}V(\mathbf{R})$ —can be obtained from $V_{\text{ét},2}$ with its involution. This question is answered affirmatively by

THEOREM 2. *Let p be a prime, π a finite p -group, X a finite-dimensional π -CW complex with finite-dimensional mod p homology. Then the composite*

$$\begin{array}{ccc} \mathbf{Z}_{p\infty}(X^\pi) & \rightarrow & (\mathbf{Z}_{p\infty}X)^{h\pi} \\ & \searrow & \nearrow \\ & (\mathbf{Z}_{p\infty}X)^\pi & \end{array}$$

is a weak homotopy equivalence.

Theorem 2 has been proven by Lannes [22], the author ([33], using joint work [13] with Neisendorfer and W. G. Dwyer), and by Carlsson [44]. The special case in which $X^\pi = \emptyset$ is due to E. Dror Farjoun and A. Zabrodsky [15] and to S. Jackowsky [26].

Theorem 2 differs somewhat from (SC). This is because (1) Sullivan followed [4] in regarding the space $V_{\text{ét},2}$ as defined only up to homotopy. The “rigidification” necessary to obtain a space, functorially, was carried out in essence by S. Lubkin in 1967, and in greater generality by E. M. Friedlander (see [17]).

(2) His description of p -adic completion also yields only a homotopy-type, functorially. We make some further remarks about these two statements.

REMARK 1. Another motivation for the conjecture comes from P. A. Smith theory. Let π be a finite p -group, let X and Y be π -spaces to which Theorem 2 applies, and let $f: X \rightarrow Y$ be an equivariant mod p homology equivalence. By (BK3), $\mathbf{Z}_{p\infty}X \rightarrow \mathbf{Z}_{p\infty}Y$ is a homotopy equivalence. By (H), then, the bottom arrow in the diagram

$$\begin{array}{ccc} \mathbf{Z}_{p\infty}(X^\pi) & \rightarrow & \mathbf{Z}_{p\infty}(Y^\pi) \\ \downarrow & & \downarrow \\ (\mathbf{Z}_{p\infty}X)^{h\pi} & \rightarrow & (\mathbf{Z}_{p\infty}Y)^{h\pi} \end{array}$$

is an equivalence, and by Theorem 2 the vertical arrows are too. By (BK3) again, $X^\pi \rightarrow Y^\pi$ is thus a mod p homology isomorphism: this is the conclusion of Smith theory. (SC), on the other hand, does not recover Smith theory.

REMARK 2. In Theorem 2 it suffices to assume that $|\pi| = p$; the general case follows by an easy induction on $|\pi|$ using the solvability of p -groups. (SC) cannot be so reduced.

REMARK 3. Obstruction theory gives

DWYER'S LEMMA [12]. *If π is a finite p -group and X is a simply connected π -space, then $X^{h\pi} \rightarrow (\mathbf{Z}_{p\infty}X)^{h\pi}$ is a mod p homology equivalence.*

This lemma enables us to compare Theorem 2 with (SC), by considering the commutative square (in which X is as in Theorem 2)

$$\begin{array}{ccc} X^\pi & \rightarrow & X^{h\pi} \\ \downarrow & & \downarrow \\ \mathbf{z}_{p\infty}(X^\pi) & \rightarrow & (\mathbf{Z}_{p\infty}X)^{h\pi} \end{array} \quad (*)$$

If X^π is simply connected, then by (BK2) the left vertical is a mod p homology equivalence; if also X is simply connected then by Dwyer's Lemma the right vertical is too; so if both conditions hold, then (SC) is true. On the other hand, according to A. K. Bousfield, there are connected finite complexes W such that $W \rightarrow \mathbf{Z}_{p\infty}W$ is not a mod p homology equivalence—for example, $W = S^1 \vee S^1$. Consider, for such W , the join $X = W * \mathbf{Z}_p$, with \mathbf{Z}_p acting trivially on W and by translation on \mathbf{Z}_p . X is simply connected, and so Theorem 2 and Dwyer's Lemma show that (SC) fails for this example.

2. Methods and extensions. The basic approach to these problems is via the technology of Bousfield and D. M. Kan [6] underlying the construction of their \mathbf{Z}_p -completion and the associated unstable Adams spectral sequence. This approach, initiated in [31], was inspired by W. H. Lin's proof [27] of the conjecture of M. E. Mahowald and G. B. Segal about the *stable* groups $\pi_S^*(\mathbf{R}P\infty)$, and was brought to life by Carlsson's paper [8] on the Segal conjecture for \mathbf{Z}_2^r . This technology results in an obstruction theory, in which obstructions lie in suitable Ext groups. These groups are certain nonabelian derived functors, computed in the category \mathcal{CA} of right unstable coalgebras over the mod p Steenrod

algebra \mathcal{A} —or, under appropriate finiteness assumptions, in the category \mathcal{K} of left unstable algebras over \mathcal{A} . For example, [6] and [11] yield the

CONTRACTIBILITY CRITERION (CC) [31]. Let \overline{H}_*W be p -torsion and let X be a simply connected (or merely nilpotent) space. If

$$\mathrm{Ext}_{\mathcal{CA}}^s(H_*(\Sigma^n W), H_*(X)) = 0 \quad \text{for all } n \geq s \geq 0,$$

then $\mathrm{map}_*(W, X)$ is weakly contractible.

Various stragatems for computing these Ext groups have been devised. In case $W = B\mathbf{Z}_p^r$, we have the following surprising and fundamental fact.

INJECTIVITY THEOREM (I). $H^*(B\mathbf{Z}_p^r)$ is an injective object in the category \mathcal{U} of left unstable \mathcal{A} -modules.

This result was proven (for $p = 2$) by Carlsson [8]. Carlsson's proof was extended (for $r = 1$) to $p > 2$ and given an interpretation of this form (but using homology) in [31]. The result as stated was given first in [26] by Lannes and S. Zarati. They have strengthened (I) in important ways as well; for example [25], $H^*(B\mathbf{Z}_p^r)$ is in fact injective in \mathcal{K} . Lannes and Schwartz [24] classify injectives in \mathcal{U} , and as a corollary obtain a converse to Theorem 1. See also [10] for related work.

The Injectivity Theorem can be used (see [31], or, for a quicker and somewhat more general proof, Lannes and Schwartz [23]) to obtain

THEOREM [31]. $\mathrm{Ext}_{\mathcal{CA}}^s(H_*(\Sigma^n B\mathbf{Z}_p^r), M) = 0$ for all $r \geq 0$, and $n \geq s \geq 0$, provided M is bounded above.

Together with (CC), this result yields Theorem 1 in case $W = B\mathbf{Z}_p$ and X is simply connected; this is the key case. We refer the reader to [30] or [31] for an account of the path to the general result.

We conclude this section by mentioning several recent related results of Lannes.

THEOREM 3 (LANNES, [22]). Let X be a simply connected space such that $H_*(X; \mathbf{Z}_p)$ is of finite type. Evaluation of cohomology

$$[B\mathbf{Z}_p^r, X] \rightarrow \mathrm{Hom}_{\mathcal{K}}(H^*(X), H^*(B\mathbf{Z}_p^r))$$

is then a bijection.

This result was conjectured in [32] and proven there under very special conditions.

Second, Lannes gives a workable Ext criterion for showing that a map $Z \rightarrow \mathrm{map}(B\mathbf{Z}_p^r, Y)$ is an equivalence. He uses this result for a variety of purposes: (1) Taking $Y = E\pi \times_\pi X$, with $|\pi| = p$, he obtains a proof of Theorem 2. (2) Taking $Y = \bigcup_n \Omega^n S^{n+k}$, he obtains a proof of the strong form of the Segal conjecture for p -tori, a theorem due originally to J. F. Adams, J. H. C. Gunawardena, and the author [3]. This proof represents a completion of Carlsson's original work [8] on the Segal conjecture, in which the Injectivity Theorem (somewhat disguised) first appeared.

Forms of this theorem have been proven also by Zabrodsky and by Lannes. Dwyer's proof may be sketched as follows. Lannes's Theorem 3, combined with Quillen's paper [36], implies that α is bijective for π a p -torus. (This observation is due to Adams, to the author and Wilkerson, and to Lannes [22].) On the other hand, using the well-known fact that $\text{map}(S^1, BG) \rightarrow BG$ is fiber-homotopy equivalent to $EG \times_G G_c \rightarrow BG$ ($G_c = G$ with G acting by conjugation) one sees that given $\rho: \pi \rightarrow G$,

$$\Omega \text{map}(B\pi, BG)_{B\rho} \simeq G_{c,\rho}^{h\pi}$$

($G_{c,\rho} = G$ with π acting by conjugation through ρ). Since $G_{c,\rho}^\pi = \text{Aut}(\rho)$, Theorem 2 (supplemented i.a. by Dwyer's Lemma) leads to good information about β . Finally, an induction using "nonabelian cohomology" gives the result for general π .

This result provides a powerful tool with which to attack problems about maps between classifying spaces. We give two sample applications. Sullivan [39] constructed a map $\psi^k: BSU_2 \rightarrow BSU_2$ such that $H^{2n}(\psi^k; Q)$ is multiplication by k^n , provided $k = 0$ or k is odd. Friedlander [16] and Wilkerson [40] extended this construction to obtain ψ^k on BG for any connected compact Lie group G , provided $(|W|, k) = 1$, where W is the Weyl group of G . Conversely:

THEOREM (K. ISHIGURO [19]). *If ψ^k exists on BG , then $(|W|, k) = 1$.*

THEOREM. (ZABRODSKY [42] for $k = 0$; MISLIN [34] in general). *For given k , any two maps ψ^k on BSU_2 are homotopic.*

The cohomology of BG is often free as a commutative algebra. In the remainder of this paper we take $p > 2$. Recall that Borel proved that if G is a connected compact Lie group and $(|W|, p) = 1$, then $H^*(BG; \mathbb{Z}_p)$ is polynomial.

A longstanding program (of N. Steenrod, D. Rector, C. W. Wilkerson, ...) asks to analyze simply connected spaces with polynomial cohomology (or more generally spaces with p -adically finite loop spaces) by "Lie theoretic" means: to construct a "maximal torus" and obtain analogues of the above theorem of Borel, for instance. Suppose for example that $H^*(X) = \mathbb{Z}_p[x]$, $|x| = 2n$. Using the fact that P^m is decomposable unless m is a power of p one finds $n = kp^e$ for some divisor k of $(p-1)$, and that $H^*(X) = \mathbb{Z}_p[y]^{\mathbb{Z}_k}$ where $|y| = 2p^e$ and $\mathbb{Z}_k \subset \mathbb{Z}_p^\times$. The question of possible values of e is precisely the Hopf invariant one mod p problem, and by [28] $e = 0: H^*(X) = H^*(BT_1)^{\mathbb{Z}_k}$, where T_r is the r -torus.

On the other hand, in [39] Sullivan noticed that $\mathbb{Z}_i \leq \mathbb{Z}_p^\times$ lifts to a subgroup $\mathbb{Z}_k \leq (\mathbb{Z}_p^\wedge)^\times$ and hence acts on $K(\mathbb{Z}_p^\wedge, 2) = BT_1^\wedge$. An easy Serre spectral sequence argument then shows that $E\mathbb{Z}_k \times_{\mathbb{Z}_k} BT_1^\wedge$ has cohomology $H^*(BT_1)^{\mathbb{Z}_k}$. Applying \mathbb{Z}_{p^∞} to this space gives a simply connected \mathbb{Z}_p -complete space X with this cohomology.

Joint work of Dwyer, Wilerson, and the author generalizes this story to higher rank, and provides a uniqueness statement; for example, the homotopy type of

$\mathbf{Z}_{p\infty}(E\mathbf{Z}_k \times_{\mathbf{Z}_k} BT_1^\wedge)$ is the only one fitting the above description. The appropriate algebra is largely due to Adams and Wilkerson [2]. They start with an unstable \mathcal{A} -algebra B , assumed to be a Noetherian integral domain. They show that if the fraction field F of B has transcendence degree r , then B embeds as a sub \mathcal{A} -algebra of $H^*(BT_r)$. Let W be the group of \mathcal{A} -algebra automorphism of $H^*(BT_r)$ which fix B ; $W \leq \mathrm{GL}(r, \mathbf{Z}_p)$. If B is integrally closed in F , then (by an unpublished result of Wilkerson; cf. [35]) there are generators t_1, \dots, t_n of $H^2(BT_r)$ and integers $e_j \geq 0$ such that $S = \mathbf{Z}_p[t_1^{p^{e_1}}, \dots, t_r^{p^{e_r}}]$ is stable under W and $B = S^W$.

THEOREM 5 (DWYER, MILLER, WILKERSON [14]). *Let X be a simply connected p -complete space such that $H^*(X) = B$ is an integrally closed Noetherian integral domain of transcendence degree r . Then*

- (1) $S = H^*(BT_r) : B = H^*(BT_r)^W$.
- (2) There is a "maximal torus": a map $i: BT_r^\wedge \rightarrow X$ and a W -action on BT_r^\wedge inducing the correct action in cohomology and for which i is equivariant. In particular, W lifts to a subgroup of $\mathrm{GL}(r, \mathbf{Z}_p^\wedge)$.
- (3) If $(|W|, p) = 1$, the induced map

$$\mathbf{Z}_{p\infty}(EW \times_W BT_r^\wedge) \rightarrow X.$$

is a homotopy equivalence.

SKETCH OF PROOF. By Lannes's Theorem 3, $H^*(X) \rightarrow H^*(BT_r) \rightarrow H^*(B\mathbf{Z}_p^r)$ is realized by a map $i: B\mathbf{Z}_p^r \rightarrow X$. Lannes's work with Bousfield lets us compute that $H^*(\mathrm{map}(B\mathbf{Z}_p^r, X)_i) \cong S$. By Hopf invariant one mod p , this forces each $e_j = 0$, proving (1). Thus $\mathrm{map}(B\mathbf{Z}_p^r, X)_i \cong BT_r^\wedge$, and evaluation gives a map $BT_r^\wedge \rightarrow X$ factoring i . The action of W on $B\mathbf{Z}_p^r$ induces an action on $\mathrm{map}(B\mathbf{Z}_p^r, X)$ which (again by Theorem 3) carries the component of i to itself, and this proves (2). Finally, a Serre spectral sequence argument gives (3).

The dimensions of the generators of polynomial algebras arising in this way have been described by A. Clark and J. Ewing [43]. The results presented above lead one to hope that a complete classification of p -complex simply connected spaces with Noetherian polynomial cohomology may be within reach.

REFERENCES

1. J. F. Adams, *Maps between classifying spaces. II*, Invent. Math. **49** (1978), 1-65.
2. J. F. Adams and C. W. Wilkerson, *Finite H -spaces and algebras over the Steenrod algebra*, Ann. of Math. (2) **111** (1980), 95-143, and **113** (1981), 621-622.
3. J. F. Adams, J. H. C. Gunawardena, and H. R. Miller, *The Segal conjecture for elementary abelian p -groups*, Topology **24** (1985), 435-460.
4. M. Artin and B. Mazur, *Etale homotopy*, Lecture Notes in Math., vol. 100, Springer-Verlag, 1969.
5. A. Assadi, *On the Borel-Quillen localization and Sullivan's fixed point conjecture*, Preprint, May 1986.
6. A. K. Bousfield and D. M. Kan, *Homotopy limits, completions, and localizations*, Lecture Notes in Math., vol. 304, Spinger-Verlag, 1972.
7. A. K. Bousfield, *The localization of spaces with respect to homology*, Topology **14** (1975), 133-150.

8. G. Carlsson, *G. B. Segal's Burnside ring conjecture for $(\mathbb{Z}/2)^k$* , *Topology* **22** (1983), 83–103.
9. —, *The Segal's Burnside ring conjecture and the homotopy limit problem*, *Proc. Internat. Congr. Math.* (Berkeley, 1986).
10. D. M. Davis, *A family of unstable Steenrod-modules which includes those of G. Carlsson*, *J. Pure Appl. Algebra* **35** (1985), 253–267.
11. E. Dror Farjoun, W. G. Dwyer, and D. M. Kan, *An arithmetic square for virtually nilpotent spaces*, *Illinois J. Math.* **21** (1977), 242–254.
12. W. G. Dwyer, *Maps between classifying spaces*, Preprint, December 1985.
13. W. G. Dwyer, H. R. Miller, and J. A. Neisendorfer, *Fiberwise localization and unstable Adams spectral sequences* (in preparation).
14. W. G. Dwyer, H. R. Miller, and C. W. Wilkerson (in preparation).
15. E. Dror Farjoun and A. Zabrodsky, *Fixed points and homotopy fixed points*, Preprint, May 1986.
16. E. M. Friedlander, *Exceptional isogenies and the classifying spaces of simple Lie groups*, *Ann. of Math.* (2) **101** (1975), 510–520.
17. —, *Etale homotopy theory of simplicial schemes*, *Ann. of Math. Studies*, vol. 104, Princeton Univ. Press, Princeton, N.J., 1982.
18. E. M. Friedlander and G. Mislin, *Locally finite approximation of Lie groups. I*, *Invent. Math.* **83** (1986), 425–436.
19. K. Ishiguro, *Unstable Adams operations on classifying spaces*, Preprint, May 1986.
20. S. Jackowsky, *A fixed point theorem for p -group actions*, Preprint, February 1986.
21. L. Jones, *The converse to the fixed point theorem of P. A. Smith. I*, *Ann. of Math.* (2) **94** (1971), 52–68.
22. J. Lannes, *Sur la cohomologie modulo p des p -groupes abeliens elementaires*, Preprint, February 1986.
23. J. Lannes and L. Schwartz, *A propos de conjectures de Serre et Sullivan*, *Invent. Math.* **83** (1986), 593–603.
24. —, *Sur la structure des A -modules instables injectifs*, manuscript, December 1985.
25. J. Lannes and S. Zarati, *Sur les dérivés de la déstabilisation*, Preprint, September 1983.
26. —, *Sur les \mathcal{U} -injectifs*, Preprint, January 1985.
27. W. H. Lin, *On conjectures of Mahowald, Segal, and Sullivan*, *Math. Proc. Cambridge Philos. Soc.* **87** (1980), 449–458.
28. A. Liulevicius, *The factorization of cyclic reduced powers by secondary cohomology operations*, *Mem. Amer. Math. Soc. No. 42* (1962).
29. C. McGibbon and J. A. Neisendorfer, *On the homotopy groups of a finite dimensional space*, *Comment. Math. Helv.* **59** (1984), 253–257.
30. H. R. Miller, *The Sullivan conjecture*, *Bull. Amer. Math. Soc.* **9** (1983), 75–78.
31. —, *The Sullivan conjecture on maps from classifying spaces*, *Ann. of Math.* **120** (1984), 39–87, and **121** (1985), 605–609.
32. —, *Massey-Peterson towers and maps from classifying spaces*, *Algebraic Topology (Aarhus, 1982)*, *Lecture Notes in math.*, vol. 1051, Springer-Verlag, 1984, pp. 401–417.
33. —, (in preparation).
34. G. Mislin, Preprint.
35. S. A. Mitchell and R. E. Stong, *An adjoint representation for polynomial algebras*, Preprint, April 1986.
36. D. G. Quillen, *The spectrum of an equivariant cohomology ring. I, II*, *Ann. of Math.* (2) **94** (1971), 549–602.
37. F. Quinn, *Finite nilpotent group actions on finite complexes*, *Geometric Applications of Homotopy Theory. I*, *Lecture Notes in Math.*, vol. 657, Springer-Verlag, 1978, pp. 375–407.
38. L. Schwartz, *La conjecture de Sullivan (d'après H. Miller)*, *Seminaire Bourbaki*, Exp. No. 638, November 1984.

- 39. D. Sullivan, *Geometric topology. Part I: Localization, periodicity, and Galois symmetry*, Massachusetts Inst. of Technology, Cambridge, Mass., 1970.
- 40. C. W. Wilkerson, *Self-maps of classifying spaces*, Localization in group theory and homotopy theory and related topics, Lecture Notes in Math., vol. 418, Springer-Verlag, 1974, 150–157.
- 41. A. Zabrodsky, *On phantom maps and a theorem of H. Miller*, Preprint.
- 42. —, *Maps between classifying spaces*, Preprint, February 1984.
- 43. A. Clark and J. Ewing, *The realization of polynomial algebras as cohomology rings*, Pacific J. Math. **50** (1974), 425–434.
- 44. G. Carlsson, *Equivariant stable homotopy and Sullivan's conjecture*, Preprint, December 1986.
- 45. C. A. McGibbon and J. A. Neisendorfer, *Various applications of Haynes Miller's theorem*, Conference on Algebraic Topology in Honor of Peter Hilton, Contemp. Math., vol. 37, Amer. Math. Soc., Providence, R.I., 1985.

UNIVERSITY OF NOTRE DAME, NOTRE DAME, INDIANA 46556, USA

Trees and Hyperbolic Geometry

JOHN W. MORGAN

Introduction. Let \mathbf{H}^n be the subspace

$$\{(x_0, \dots, x_n) \in \mathbf{R}^{n+1} \mid q(x_0, \dots, x_n) = -x_0^2 + x_1^2 + \dots + x_n^2 = -1\}$$

and $x_0 > 0$. The quadratic form q restricts to give a positive definite form on each tangent space of \mathbf{H}^n , and consequently, endows \mathbf{H}^n with a riemannian metric. We call this riemannian manifold *hyperbolic n -space*. It has constant sectional curvature -1 and is homogeneous. Its group of orientation-preserving isometries is $\mathrm{SO}^\circ(n, 1)$, the connected component of the identity of the real linear subgroup $\mathrm{SO}(n, 1)$ of matrices in $\mathrm{SL}_{n+1}(\mathbf{R})$ preserving the form q . A *hyperbolic n -manifold* is a complete, oriented riemannian manifold locally isometric to \mathbf{H}^n . If N is a hyperbolic n -manifold, then N is isometric to \mathbf{H}^n/Γ_N , where Γ_N is some discrete, torsion-free subgroup of $\mathrm{SO}^\circ(n, 1)$. The group Γ_N is determined by N up to conjugation in $\mathrm{SO}^\circ(n, 1)$.

Now let us fix a torsion-free, finitely generated group Γ . An *n -dimensional hyperbolic structure* on Γ is determined by a pair $\{N, \varphi\}$, where N is an n -dimensional hyperbolic manifold and $\varphi: \Gamma \rightarrow \pi_1(N)$ is an isomorphism. Two pairs $\{N_0, \varphi_0\}$ and $\{N_1, \varphi_1\}$ determine the same hyperbolic structure on Γ if there is an orientation-preserving isometry $I: N_0 \rightarrow N_1$ such that $I_* \circ \varphi_0$ and φ_1 differ by an inner automorphism of Γ . (Notice that $I_*: \pi_1(N_0) \rightarrow \pi_1(N_1)$ is only well defined up to inner automorphism of $\pi_1(N_0)$.)

We denote the set of all n -dimensional hyperbolic structures on Γ by $\mathcal{H}^n(\Gamma)$. This set is naturally identified with the set of conjugacy classes of faithful representations $\rho: \Gamma \rightarrow \mathrm{SO}^\circ(n, 1)$ with discrete image. From this description, there is a natural topology on $\mathcal{H}^n(\Gamma)$. Let $R_+^n(\Gamma)$ be the space of all representations of Γ into $\mathrm{SO}^\circ(n, 1)$. It has the algebraic topology: representations are close if they are close on a finite generating set. Inside $R_+^n(\Gamma)$ we have the subspace $\tilde{D}_+^n(\Gamma)$ of faithful representations with discrete image. $\mathrm{SO}^\circ(n, 1)$ acts on $R_+^n(\Gamma)$ by conjugation leaving $\tilde{D}_+^n(\Gamma)$ invariant. The quotient space $D_+^n(\Gamma) = \tilde{D}_+^n(\Gamma)/\mathrm{SO}^\circ(n, 1)$ is the space of conjugacy classes of discrete, faithful representations of Γ into $\mathrm{SO}^\circ(n, 1)$. Since we have a natural identification $\mathcal{H}^n(\Gamma) \leftrightarrow D_+^n(\Gamma)$, this yields a

topology on $\mathcal{H}^n(\Gamma)$. As an example, if Γ is the fundamental group of a closed surface of genus $g > 1$, then $\mathcal{H}^2(\Gamma)$ is the Teichmüller space of Γ . It is homeomorphic to \mathbf{R}^{6g-6} .

The basic object of study here is the space $\mathcal{H}^n(\Gamma)$. We are particularly concerned with degenerations of hyperbolic structures, i.e., unbounded sequences in $\mathcal{H}^n(\Gamma)$. We address this by constructing a natural compactification of $\mathcal{H}^n(\Gamma)$ and giving a geometric realization of the ideal points of this compactification. One question which naturally arises is, What conditions on Γ ensure that the space $\mathcal{H}^n(\Gamma)$ is compact? One class of groups Γ for which $\mathcal{H}^n(\Gamma)$ has been explicitly determined is the class of fundamental groups of finite volume hyperbolic n -manifolds N . Mostow rigidity [9] says that for $n > 2$, $\mathcal{H}^n(\pi_1(N))$ is a single point. We view results saying that $\mathcal{H}^n(\Gamma)$ is compact for a wider class of groups Γ as a weak version of Mostow rigidity for infinite volume hyperbolic manifolds. Of course, if Γ is a surface group $\mathcal{H}^2(\Gamma) \subset \mathcal{H}^n(\Gamma)$ for all $n \geq 2$. Thus there can be no general result saying that $\mathcal{H}^n(\Gamma)$ is compact for all finitely generated Γ . What is more surprising is that there are large classes of groups for which $\mathcal{H}^n(\Gamma)$ is nonempty (and even positive-dimensional) but still compact. The full story is not yet understood but some significant partial results do exist.

Results. The first such result (beyond Mostow rigidity) is due to Thurston.

THEOREM 1 [11]. *Let M^3 be a hyperbolic 3-manifold and suppose that $\Gamma = \pi_1(M)$ is finitely generated. If $\mathcal{H}^3(\Gamma)$ is noncompact, then Γ has a decomposition of one of the following types:*

- (i) $\Gamma \cong A *_C B$ is a nontrivial free product with amalgamation with C cyclic;
- (ii) $\Gamma \cong A *_C$ is an HNN-decomposition with C cyclic.

In the proof of this result, the fact that Γ is a 3-manifold group is important but the dimension of the hyperbolic structure is much less important. In fact, one shows

THEOREM 2 [4]. *Let M be a hyperbolic 3-manifold, and suppose that $\Gamma = \pi_1(M)$ is finitely generated. Then for any $n \geq 3$, $\mathcal{H}^n(\Gamma)$ is compact if and only if Γ does not have a decomposition of one of the following types:*

- (i) $\Gamma \cong A *_C B$ with C cyclic and of infinite index in A and B ;
- (ii) $\Gamma \cong A *_C$ with C cyclic.

Notice that it follows that $\mathcal{H}^n(\pi_1(M))$ is compact if and only if $\mathcal{H}^3(\pi_1(M))$ is compact. In particular, if M is a finite volume hyperbolic 3-manifold, then $\mathcal{H}^n(\pi_1(M))$ is compact for all $n \geq 3$.

There is another class of groups for which very similar results hold. We state the result in the case of no parabolic subgroups.

THEOREM 3. *Let Γ be a finitely generated group. Suppose all abelian subgroups of Γ are cyclic. Let $I(\Gamma)$ be the set of subgroups $\Gamma' \subset \Gamma$ of infinite index. Suppose there is a class $\alpha \in H_k(\Gamma; \mathbf{Q})$, for some $k \geq 3$, which is not in the image*

of the natural map

$$\bigoplus_{\Gamma' \in I(\Gamma)} H_k(\Gamma'; \mathbf{Q}) \rightarrow H_k(\Gamma; \mathbf{Q}). \quad (*)$$

Then $\mathcal{H}^n(\Gamma)$ is compact provided that Γ does not admit a decomposition of either of the following forms:

- (I) $\Gamma = A *_C B$ with C virtually abelian and $C \neq A$, $C \neq B$;
- (II) $\Gamma = A *_C$ with C virtually abelian.

There is also a version of this theorem in the presence of parabolic subgroups. One uses homology relative to the cusps.

One class of groups for which this result gives information is groups Γ for which the Eilenberg-Mac Lane space $K(\Gamma, 1)$ has a representative which is a closed, oriented k -manifold, $k \geq 3$. Then Γ has a top class $[\Gamma] \in H_k(\Gamma; \mathbf{Q})$ which is not in the image of $(*)$. Thus we have

COROLLARY 4. *Suppose that M^k , $k \geq 3$, is a closed oriented k -manifold with $\pi_1(M) \cong \Gamma$ and universal covering \tilde{M} contractible. Then for any n , $\mathcal{H}^n(\Gamma)$ is compact provided that Γ has no decomposition as in (I) and (II) above.*

In particular, we have

COROLLARY 5. *Suppose that M^k , $k \geq 3$, is a closed, hyperbolic k -manifold with $\pi_1(M) \cong \Gamma$. Then for any n , $\mathcal{H}^n(\Gamma)$ is compact.*

This last result is a weak version for rank-1 groups of Margulis's super rigidity for co-compact lattices in higher rank groups [3].

There is also the version of the previous result for finite volume hyperbolic manifolds.

These results suggest the following question:

Let Γ be a finitely generated group, and let $n \geq 3$. Is $\mathcal{H}^n(\Gamma)$ compact unless Γ has a decomposition as in (I) or (II) of Theorem 3?

Sketch of the proofs. We shall give a rough sketch of the ideas that go into the proofs of these results. We start by describing the space $A(\Gamma)$ of actions of Γ on \mathbf{R} -trees, which is the analogue of the character variety of representations of Γ into $\mathrm{SO}(n, 1)$. Then we go into some of the details of the compactification of $\mathcal{H}^n(\Gamma)$ and the identification of the ideal points of the compactification with actions of Γ on \mathbf{R} -trees. After this Theorems 1 and 2 are direct consequences of a theorem due to Shalen and myself on 3-manifold groups acting on \mathbf{R} -trees. We also briefly discuss the proof of Theorem 3.

Actions on \mathbf{R} -trees. The study of degenerations of hyperbolic structures on Γ is closely related to actions of Γ on \mathbf{R} -trees. Our next goal is to study actions of Γ on \mathbf{R} -trees in their own right.

An \mathbf{R} -tree is a metric space (T, d) such that (i) any two points x and y in T are joined by a segment¹ in T and (ii) if I and J are segments in T with

¹A segment in a metric space is a subspace isometric to a closed interval in \mathbf{R} .

$I \cap J = \{x\}$, x being an endpoint of both I and J , then $I \cup J$ is a segment. (It follows that the segment in T joining x and y is unique.) Let $\gamma: T \rightarrow T$ be an isometry. There are two possibilities:

- (a) γ has a fixed point, or
- (b) there is a copy of \mathbf{R} isometrically embedded in T which is invariant under γ and on which γ acts as a nontrivial translation.

In case (b), the image of \mathbf{R} is unique and is called the *axis* of γ . We define the *hyperbolic length* of γ to be

$$\tau(\gamma) = \min_{x \in T} d(x, \gamma x).$$

Case (a) corresponds to $\tau(\gamma) = 0$, and case (b) corresponds to $\tau(\gamma) > 0$.

Now suppose that Γ is a finitely generated group and that $\psi: \Gamma \times T \rightarrow T$ is an action of Γ on an \mathbf{R} -tree. (We use the term action to mean action by isometries.) We define the *hyperbolic length function* $\tau_\psi: \Gamma \rightarrow \mathbf{R}$ by $\tau_\psi(\gamma) = \tau(\psi(\gamma))$. Since τ_ψ is a class function, $\tau_\psi(\gamma_0 \gamma_1 \gamma_0^{-1}) = \tau_\psi(\gamma_1)$, it is natural to factor τ_ψ as a function $\tau_\psi: \mathcal{C} \rightarrow \mathbf{R}$, where \mathcal{C} is the set of conjugacy classes in Γ . In this way an action gives rise to a point in $[0, \infty)^{\mathcal{C}}$. A hyperbolic length function is said to be *abelian* if it is the absolute value of a homomorphism $\Gamma \rightarrow \mathbf{R}$. The first key lemma is from [2].

LEMMA 6. (a) *An action $\Gamma \times T \rightarrow T$ has a fixed point, i.e., there is $x \in T$ with $\gamma x = x$ for all $\gamma \in \Gamma$, if and only if the hyperbolic length function is zero.*

(b) *If the hyperbolic length function of the action is nonzero, then there is a unique minimal Γ -invariant subtree $T_0 \subset T$.*

(c) *Two actions of Γ with nonabelian hyperbolic length functions have Γ -equivariantly isomorphic minimal Γ -invariant subtrees if and only if their hyperbolic length functions are equal.*

Using this theorem we can form a “space” of minimal actions of Γ without fixed points. It is the subspace $\tilde{A}(\Gamma) \subset [0, \infty)^{\mathcal{C}} - \{0\}$ of all nontrivial hyperbolic length functions. It is convenient to form the projectivized version of these spaces $A(\Gamma) \subset P(\mathcal{C}) = ([0, \infty)^{\mathcal{C}} - \{0\})/\mathbf{R}^+$. We think of $A(\Gamma)$ as an analogue of the character variety of Γ .

Compactification of $\mathcal{H}^n(\Gamma)$. If $\{N, \varphi\}$ represents a point $x \in \mathcal{H}^n(\Gamma)$ and if $\gamma \in \Gamma$, then $\varphi(\gamma) \in \pi_1(N) \subset \mathrm{SO}^0(n, 1)$ is either hyperbolic, in which case its free homotopy class in N contains a unique closed geodesic, or $\varphi(\gamma) \in \pi_1(N)$ is parabolic, in which case its free homotopy class in N contains arbitrarily short loops. We define the length $l_x(\varphi(\gamma))$ to be the length of the closed geodesic in the first case and to be 0 in the second case. Provided that Γ is not virtually abelian, this yields a map $\tilde{\Theta}: \mathcal{H}^n(\Gamma) \rightarrow [0, \infty)^{\mathcal{C}} - \{0\}$ defined by $\tilde{\Theta}(x) = \{l_x(\gamma)\}_{\gamma \in \mathcal{C}}$. Once again, it is convenient to projectivize to obtain $\Theta: \mathcal{H}^n(\Gamma) \rightarrow P(\mathcal{C})$. The compactification of $\mathcal{H}^n(\Gamma)$ is derived from the following result.

LEMMA 7. *Let Γ be a finitely generated, nonvirtually abelian group. The closure of the image $\Theta(\mathcal{H}^n(\Gamma)) \subset P(\mathcal{C})$ is compact.*

This leads to a natural compactification $\overline{\mathcal{H}}^n(\Gamma)$ for $\mathcal{H}^n(\Gamma)$. The ideal points for this compactification form the subspace $B^n(\Gamma) \subset P(\mathcal{C})$ of limit points of all sequences $\{\Theta(x_i)\}$, where $\{x_i\} \in \mathcal{H}^n(\Gamma)$ is an unbounded sequence. Of course, $\overline{\mathcal{H}}^n(\Gamma) = \mathcal{H}^n(\Gamma) \cup B^n(\Gamma)$, where $B^n(\Gamma)$ is glued on at infinity in the obvious way.

In the case $n = 2$, this is exactly Thurston's compactification of the Teichmüller space.

Relation to \mathbf{R} -trees. The connection between this compactification of $\mathcal{H}^n(\Gamma)$ and actions of Γ on \mathbf{R} -trees is the following:

PROPOSITION 8. $B^n(\Gamma) \subset A(\Gamma) \subset P(\mathcal{C})$.

This means that if $\{x_k\} \in \mathcal{H}^n(\Gamma)$ is an unbounded sequence of hyperbolic structures, then, after passing to a subsequence, we can find an action of Γ on an \mathbf{R} -tree without fixed point whose hyperbolic length function τ is such that

$$\lim_{k \rightarrow \infty} \{l_{x_k}(\gamma)\}_{\gamma \in \mathcal{C}} = \{\tau(\gamma)\}_{\gamma \in \mathcal{C}} \quad \text{in } P(\mathcal{C}).$$

If we take the action on the \mathbf{R} -tree to be minimal, then it is unique up to equivariant isometry. The reason is that since Γ is not virtually abelian, the limit hyperbolic length function can never be abelian. (Actually, much more is true; see Addendum 8' below.)

Once again, this is the analogue of Thurston's result that the ideal points in the compactification of Teichmüller space are interpreted as measured geodesic laminations on the surface [12]. The connection between the two results is explained in [6], where it is shown that the actions occurring in $B^2(\Gamma) \subset A(\Gamma)$ are in fact dual to measured laminations on the closed surface with fundamental group Γ .

Actually, there is an analogous compactification of the entire space of conjugacy classes of representations of Γ into $\mathrm{SO}^\circ(n, 1)$. In this case as well the ideal boundary is a subspace of $A(\Gamma)$. There is, however, something special about the ideal boundary of $\mathcal{H}^n(\Gamma)$, and this is a consequence of the Margulis lemma about short loops in hyperbolic manifolds.

We say that a group H is *small* if it contains no free group of rank 2. If $\mathcal{H}^n(\Gamma) \neq \emptyset$, for some n , then any small subgroup of Γ is in fact virtually abelian.

ADDENDUM 8'. *Let Γ be a finitely generated, nonvirtually abelian group. Then $B^n(\Gamma)$ is contained in the subspace of $A(\Gamma)$ represented by actions $\Gamma \times T \rightarrow T$ with the property that every nondegenerate segment in T has stabilizer in Γ which is small (i.e., virtually abelian).*

One can establish the result that the ideal points of the compactification of $\mathcal{H}^n(\Gamma)$ are actions of Γ on \mathbf{R} -trees, i.e., that as hyperbolic manifolds degenerate they degenerate to actions on \mathbf{R} -trees, either algebraically or geometrically.

Thurston [11] has a method of constructing the \mathbf{R} -tree by looking at the way ideal simplices in hyperbolic space can degenerate. Gromov has a more direct method of seeing hyperbolic space itself degenerating to a universal \mathbf{R} -tree. The approach we describe here is more algebraic. There are advantages

to each approach. The geometric ones apply *mutatis mutandis* to the class of manifolds of universally bounded negative curvature, whereas the algebraic one requires homogeneity. On the other hand, the algebraic one should generalize more directly to the homogeneous spaces of higher rank Lie groups.

Let $R(\Gamma, \mathrm{SO}(n, 1))$ be the affine variety (defined over \mathbf{Q}) of representations of Γ into $\mathrm{SO}(n, 1)$. Let $X(\Gamma, \mathrm{SO}(n, 1))$ be the quotient, in the category of affine varieties, of $R(\Gamma, \mathrm{SO}(n, 1))$ by the conjugation action of $\mathrm{SO}(n, 1)$. Then $D_+^n(\Gamma)$ maps by a proper, finite-to-one map to $X^n(\Gamma) = X_{\mathbf{R}}(\Gamma, \mathrm{SO}(n, 1))$. Let $D^n(\Gamma)$ be the image. We compactify $X^n(\Gamma)$ as follows: We have a map $\Theta: X^n(\Gamma) \rightarrow P(\mathcal{C})$ which associates to the class of a representation ρ the point $\{\ln(|\mathrm{tr} \rho(\gamma)| + 2)\}_{\gamma \in \mathcal{C}}$. According to [7, §1] and Lemma 3.1 of [4] $\mathrm{Im} \theta$ has compact closure in $P(\mathcal{C})$. Let $B(X^n(\Gamma)) \subset P(\mathcal{C})$ be the set of ideal points. Inside $X^n(\Gamma)$ we have the closed subspace $Y^n(\Gamma)$ which is the image of the real representations. This leads to a compactification of $Y^n(\Gamma)$ with space of ideal points $B(Y^n(\Gamma)) \subset B(X^n(\Gamma))$. We have (see Theorem 3.2 of [4])

FACT 9. *If $b \in B(Y^n(\Gamma)) \subset P(\mathcal{C})$, then there is a \mathbf{Q} -subvariety $W_b \subset R(\Gamma, \mathrm{SO}(n, 1))$ and a valuation*

$$v_b: \mathbf{Q}(W_b)^\times \rightarrow \Lambda_b$$

such that

- (i) *the residue field of v_b is formally real and*
- (ii) $b = \{\max(0, -v_b(\mathrm{tr}_\gamma|W_b))\}_{\gamma \in \mathcal{C}}$.

Explanation of (ii). $\mathrm{tr}_\gamma|W_b$ is a polynomial function on W_b , i.e., an element of $\mathbf{Q}[W_b] \subset \mathbf{Q}(W_b)$. Thus, the term $m_\gamma = \max(0, -v_b(\mathrm{tr}_\gamma|W_b)) \in \Lambda_b$. Since Λ_b is not necessarily archimedean, some discussion is necessary to interpret $\{m_\gamma\}_{\gamma \in \mathcal{C}}$ as an element of $P(\mathcal{C})$. Let $\Lambda_b^1 \subset \Lambda_b$ be the convex subgroup generated by the m_γ . Let $\Lambda_b^0 \subset \Lambda_b^1$ be its maximal proper convex subgroup. Then Λ_b^1/Λ_b^0 is archimedean, and hence embeds in an order-preserving fashion in \mathbf{R} , uniquely up to positive scalar factor. In this way the m_γ determine real numbers unique up to a common scale factor. Hence, the indexed set $\{m_\gamma\}_{\gamma \in \mathcal{C}}$ determines a point in $P(\mathcal{C})$ independent of the choice of scaling.

Now comes the Bruhat-Tits local building [1]. Let K be a field and $v: K^\times \rightarrow \Lambda$ a valuation with formally real residue field. Then there is an \mathbf{R} -tree T on which $\mathrm{SO}_K(n, 1)$ acts naturally; the hyperbolic length function of the action is $\tau(\alpha) = \max(0, -v(\mathrm{tr}_\alpha))$ for all $\alpha \in \mathrm{SO}_K(n, 1)$. These elements of Λ are interpreted, as before, in \mathbf{R} . Since we have the tautological representation Γ into $\mathrm{SO}(n, 1)$ over the function field $\mathbf{Q}(W_b)$, this yields:

COROLLARY 10. *For each point $p \in B(Y^n(\Gamma)) \subset P(\mathcal{C})$ there is an action of Γ on an \mathbf{R} -tree with hyperbolic length function τ such that $b = \{\tau(\gamma)\}_{\gamma \in \mathcal{C}}$.*

Restricting to the subspace characters of discrete and faithful representations $D^n(\Gamma) \subset Y^n(\Gamma)$ gives a compactification of it, and hence of $D_+^n(\Gamma) = \mathcal{H}^n(\Gamma)$, whose ideal points in $P(\mathcal{C})$ are hyperbolic length functions of actions of Γ on \mathbf{R} -trees. This proves Proposition 8.

Proofs of the theorems. As a consequence of Proposition 8 and Addendum 8' we have the following. We say that an action of Γ on an \mathbf{R} -tree has *small edge stabilizers* if the stabilizer of every nondegenerate segment is small.

PROPOSITION 11. *Suppose every action of Γ on an \mathbf{R} -tree T with small edge stabilizers has a fixed point. Then for n , $\mathcal{H}^n(\Gamma)$ is compact.*

Thus, one approach to the study of when $\mathcal{H}^n(\Gamma)$ is compact is the study of when all actions of Γ on \mathbf{R} -trees with small segment stabilizers have fixed points. If we replace \mathbf{R} -trees by simplicial trees, then the answer is well understood. An action of Γ on a simplicial tree with small segment stabilizers corresponds to a graph product decomposition of Γ over small subgroups. The decomposition is trivial if and only if the action has a fixed point. Thus, all actions of Γ on simplicial trees with small segment stabilizers have fixed points if and only if Γ has no decomposition as in (I) or (II) of Theorem 3. For some classes of groups the same result holds for actions on \mathbf{R} -trees.

THEOREM 12 [8]. *Let Γ be a finitely generated group which is the fundamental group of a 3-manifold. If Γ has an action without fixed points on an \mathbf{R} -tree with small edge stabilizers, then Γ has a decomposition as in (I) or (II) in Theorem 3.*

The idea is to show one can replace the action of Γ on an \mathbf{R} -tree with these properties by one of Γ on a simplicial tree with the same properties.

Theorem 1 is immediate from Proposition 11, Theorem 12, and the remark that a nontrivial decomposition of the fundamental group of a hyperbolic 3-manifold as in (I) or (II) of Theorem 3 must automatically have C cyclic. Theorem 2 also follows from the same line of reasoning. (For the converse, it is easy to construct degenerations when Γ has a decomposition as in (i) or (ii) of Theorem 2.)

Of course, Proposition 11 suggests the following question:

If Γ acts without fixed points on an \mathbf{R} -tree with the stabilizer of every nondegenerate segment being small, then is there an action of Γ on a simplicial tree with the same properties? An affirmative answer to this question would yield, via Proposition 11, an affirmative answer to the question following Corollary 5.

The proof of Theorem 3 relies on a more geometric description of the limit action of Γ on an \mathbf{R} -tree coming from a degenerating sequence of hyperbolic structures on Γ . This is achieved by following Thurston's ideas on the shapes of degenerating ideal simplices. Details can be found in [5].

History. Compactifications of hyperbolic structures in this vein began with Thurston's compactification of the Teichmüller space by adding geodesic measured laminations at infinity [12]. The first compactness result for infinite volume hyperbolic structures is Thurston's result (Theorem 1) proved in [11]. Shalen and I developed in [7] the theory, described here for $\mathrm{SO}(n, 1)$, for $\mathrm{SL}_2(\mathbf{C})$. This allowed us, once we have proved Theorem 12, to obtain Thurston's result from the point of view adopted here.

The idea that \mathbf{R} -trees are related to $\mathrm{SO}_K(n, 1)$ and valuations v on K with formally real residue field was known to Bruhat-Tits [1]. The idea of describing these buildings in terms of lattices in the case of SL_2 originated with Serre [10] who considered discrete, rank-1 valuations. Shalen and I generalized this in [7] to the case of SL_2 and arbitrary valuations.

REFERENCES

1. F. Bruhat and J. Tits, *Groupes réductifs sur un corps local*, Inst. Hautes Études Sci. Publ. Math. **41** (1972), 5–252.
2. M. Culler and J. Morgan, *Group actions on \mathbf{R} -trees*, Proc. London Math. Soc. (to appear)
3. A. Margulis, *Arithmeticity of nonuniform lattices*, J. Funct. Anal. **7** (1973), 245–246.
4. J. Morgan, *Group actions on trees and the compactification of the space of classes of $\mathrm{SO}(n, 1)$ -representations*, Topology **25** (1986), 1–33.
5. —, *Trees and hyperbolic geometry*, CBMS Regional Conf. Ser. in Math., Amer. Math. Soc., Providence, R.I. (to appear).
6. J. Morgan and J.-P. Otal, *Actions of surface groups on Λ -trees*, Preprint.
7. J. Morgan and P. Shalen, *Valuations, trees, and degenerations of hyperbolic structures. I*, Ann. of Math (2) **120** (1984), 401–476.
8. —, *Valuations, trees, and degenerations of hyperbolic structures. II* (to appear).
9. G. Mostow, *Quasi-conformal mappings in n -space and the rigidity of hyperbolic space forms*, Inst. Hautes Études Sci. Publ. Math. **34** (1968), 53–104.
10. J.-P. Serre, *Trees*, Springer-Verlag, New York, 1980.
11. W. Thurston, *Hyperbolic structures on 3-manifolds. I: Deformation of acylindrical manifolds*, Ann. of Math (2) **124** (1986), 203–246.
12. —, *On the geometry and dynamics of diffeomorphisms of surfaces. I*, photocopied research announcement, Princeton Univ., 1982.

COLUMBIA UNIVERSITY, NEW YORK, NEW YORK 10027, USA

Applications of Topology with Control

FRANK QUINN

Topology “with control” has been an important tool in the study of the topology of manifolds, with applications ranging from a topological characterization of manifolds to the 4-dimensional Annulus Conjecture. However the methods seem to find their most natural and powerful expression in the context of “singular spaces” and stratified sets. This has had interesting applications to topological actions of groups on manifolds, and holds promise for problems like the topological stability of smooth maps and the topological structure of algebraic varieties.

In this note we describe a prototypical result, the controlled h -cobordism theorem, and follow it through the development of the theory. First it is applied to obtain uniqueness results for neighborhoods in certain singular spaces. “Weakly stratified sets” are then defined, as a setting in which the controlled hypotheses can easily and naturally be verified. An h -cobordism theorem for weakly stratified sets is described. Finally we discuss the “rigidity” phenomenon. In this the differential geometry, algebra, and topology of certain groups of isometries are drawn together by controlled topology.

1. Controlled topology. Suppose W is a compact manifold with boundary $\partial W = M_0 \cup M_1$. W is an h -cobordism if it deformation retracts to both M_0 and M_1 . The classical h -cobordism theorem asserts that W is isomorphic to $M_0 \times [0, 1]$ provided an obstruction $\tau(W, M_0)$ in the Whitehead group $\text{Wh}(\pi_1 W)$ vanishes, and $\dim W \geq 6$. More precisely, isomorphism classes of h -cobordisms with end M_0 are taken bijectively to the obstruction group by the invariant τ .

Now suppose X is a metric space and $W \rightarrow X$ is given. The *controlled* h -cobordism theorem gives criteria for the existence of a product structure $W \cong M_0 \times [0, 1]$ so that the images of the arcs $\{m\} \times [0, 1]$ in X have small diameter. To make this precise suppose $\varepsilon > 0$. A homotopy $h: M \times [0, 1] \rightarrow X$ has *radius less than ε* if for every m the arc $h(\{m\} \times [0, 1])$ is contained in the ball of radius ε about $h(m, 0)$. (If X is noncompact we use control functions $\varepsilon: X \rightarrow (0, \infty)$ rather than constant ε . In this case we require that the arc be in the ball of radius $\varepsilon(h(m, 0))$.) We can now formulate the goal as a product structure on W which

Partially supported by the National Science Foundation.

has radius $< \varepsilon$ as a homotopy. The appropriate data is an (ε, h) -cobordism: a W which deformation retracts to M_0 and M_1 , by deformations whose projections into X have radius $< \varepsilon$.

As in the classical theorem there is an algebraic obstruction related to fundamental groups. In the controlled situation the fundamental groups which appear are roughly those of inverses of points in X . In order to organize these we introduce a reference map $r: E \rightarrow X$, and require that $W \rightarrow X$ factors through a $(\delta, 1)$ -connected map $W \rightarrow E$. We will want r to be “reasonable” in the sense that the point inverses do not vary too wildly. Technically, simplicial maps are “reasonable,” as are retracts of simplicial maps, and “stratified systems of fibrations with ANR filtration” [6, part II].

1.1 THEOREM [6, PART II]. *Suppose $r: E \rightarrow X$ is “reasonable,” and $n, \varepsilon > 0$ are given. Then there is a $\delta > 0$ so that if $W^n \rightarrow E$ is $(\delta, 1)$ -connected and a (δ, h) -cobordism over X , then an invariant $\tau(W, M_0)$ is defined in $H_1^{lf}(X; \mathbf{Wh}(r))$. If $n \geq 6$ this defines a bijection between ε isomorphism classes of such (δ, h) -cobordisms beginning with M_0 and the obstruction group.*

Here $\mathbf{Wh}(F)$ denotes the spectrum-valued functor with homotopy groups the Whitehead groups $\mathrm{Wh}_i(\pi_1 f)$. The obstruction group is formed (roughly) by applying $\mathbf{Wh}(-)$ fiberwise to $r: E \rightarrow X$ to obtain a “system of spectra over X .” Then form locally finite homology with coefficients in this system of spectra [6, part II]. This is a rather complicated construction, but has a great many useful formal properties.

Note that this only depends on π_1 of the fibers of r . As a consequence if $r \rightarrow s$ is a morphism of maps to X which is 1-connected on fibers it induces an isomorphism of obstruction groups. Also since the Whitehead groups of the trivial group vanish, all h -cobordisms are products if M_0 itself is $(\delta, 1)$ -connected over X .

The proof of this theorem has two parts. The first identifies the obstruction as an element of a “controlled Whitehead group” which depends on ε [6, part I; 1; 8]. There is a lot of technical detail, but this is a relatively straightforward extension of the classical proof. The second step, which is more difficult, is the identification of this controlled group with the homology group [6, part II].

2. Uniqueness of neighborhoods. We describe two uniqueness results for neighborhoods in singular spaces which use the controlled h -cobordism theorem. These also illustrate the relevance of two point-set notions: tameness and the homotopy link [10].

The *homotopy link* $\mathrm{holink}(X, Y)$ is the space of maps $f: [0, 1] \rightarrow X$ with $f^{-1}(Y) = \{0\}$. It maps to Y by evaluation at 0, and to $X - Y$ by evaluation at 1. The evaluation $p: \mathrm{holink}(X, Y) \rightarrow Y$ is a homotopy approximation to a map for a mapping cylinder neighborhood of Y in X , in the sense that if such a neighborhood does exist with map $g: N \rightarrow Y$, then there is a natural

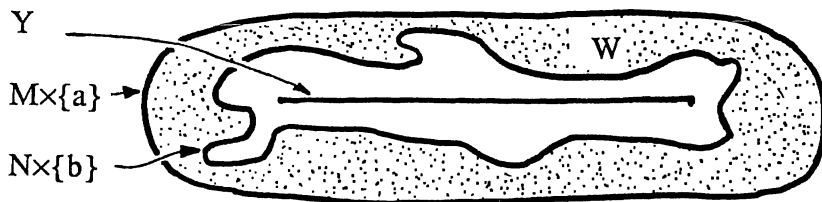


FIGURE 1

approximate fiber homotopy equivalence $g \rightarrow p$ over Y . The homotopy link is used to specify local fundamental groups of the complement near Y .

A subset $Y \subseteq X$ is *tame* if there is a deformation retraction of a neighborhood to Y , which keeps the complement in the complement until the end of the deformation. A neighborhood of a tame subset has the same local homotopy properties as a mapping cylinder neighborhood, and in particular acts like the mapping cylinder of the homotopy link.

If $Y \subseteq X$ is tame, $X - Y$ is a manifold of dimension at least 6, and the homotopy link is "reasonable," then there is an obstruction in $H_0^{lf}(Y; \mathbf{Wh}(p))$ which vanishes if and only if Y has a mapping cylinder neighborhood [6, part II; 10]. Similarly there is an obstruction in H_1^{lf} to the uniqueness of a mapping cylinder neighborhood structure. This second obstruction comes from the controlled h -cobordism theorem, as we will describe.

Suppose Y has a mapping cylinder neighborhood in X with map $f: M \rightarrow Y$. This is said to be a *manifold* mapping cylinder if M is a manifold. If we are given a second manifold mapping cylinder structure, with map $g: N \rightarrow Y$, then we can construct an h -cobordism. Choose some $a > 0$, and $b > 0$ small enough that the image of $M \times [0, a]$ in X contains the image of $N \times [0, b]$. Let W denote the closure of the region between these images (see Figure 1).

If b is small enough we can use radial deformations in the mapping cylinders to construct deformation retractions of W to the images of $M \times \{a\}$ and $N \times \{b\}$. If a is small then these deformations have small radius in Y , because the radial deformations are the identity on Y . For appropriate a we therefore get a (δ, h) -cobordism over Y . The appropriate reference map for local fundamental groups is $p: \text{holink}(X, Y) \rightarrow Y$.

Note that if this h -cobordism W has a product structure we can construct an isotopy which matches up the images of $M \times \{a\}$ and $N \times \{b\}$, and then we can match $M \times \{a\}$ and $N \times \{a\}$ by radial deformation in either mapping cylinder. We say that the mapping cylinder neighborhood structures are *isotopic* if there is an isotopy fixing Y which matches up the images of all levels $M \times \{t\}$ and $N \times \{t\}$.

2.1 THEOREM. *Suppose X, Y are as above, $p: \text{holink}(X, Y) \rightarrow Y$ is "reasonable," and $\dim(X - Y) \geq 6$. Then isotopy classes of manifold mapping cylinder neighborhoods are classified by the h -cobordism obstruction in $H_1^{lf}(Y; \mathbf{Wh}(p))$.*

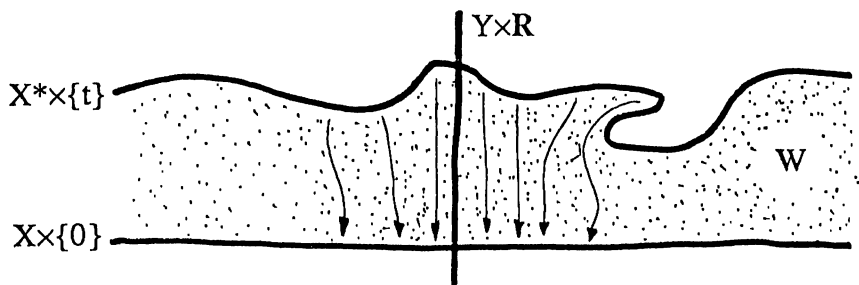


FIGURE 2

The complete proof uses the argument above to match up a sequence of levels $M \times \{a_i\}$ with $a_i \rightarrow 0$, and then the controlled pseudoisotopy result of [6, part IV] to fill in between these. There is an existence theorem for mapping cylinder neighborhoods of tame Y , with obstruction in the next lower group $H_0^{lf}(Y; \mathbf{Wh}(p))$ [6, part II]. Note that if $X - Y$ is locally 1-connected at Y then the obstruction groups vanish, so mapping cylinders exist and are unique up to isotopy. This often happens in manifolds [6, part I].

The second application of the h -cobordism theorem concerns neighborhoods of products. Suppose $Y \times \mathbf{R} \subseteq Z$. Then a pair (X, Y) and an extension of the inclusion to a homeomorphism $X \times \mathbf{R} \rightarrow Z$ onto a neighborhood of $Y \times \mathbf{R}$ will be called a *factoring* of the neighborhood. A *manifold* factoring is one with $X - Y$ a manifold. A second factoring $(X^*) \times \mathbf{R} \rightarrow Z$ is *isotopic* to the first provided there is an isotopy of Z fixing $Y \times \mathbf{R}$ which matches up the images of $(X) \times \{t\}$ and $(X^*) \times \{t\}$ near $Y \times \mathbf{R}$, for all t .

Given two manifold factorings we construct an h -cobordism which will be the obstruction to isotopy. If $t > 0$ then the images of $X \times \{0\}$ and $(X^*) \times \{t\}$ are disjoint near $Y \times \mathbf{R}$; denote by W the region between them, minus $Y \times [0, t]$ (see Figure 2). Using pushes in the \mathbf{R} coordinates we can define deformation retractions of W to the images of $(X - Y) \times \{0\}$ and $(X^* - Y) \times \{t\}$, at least near $Y \times \mathbf{R}$.

As above the pushes are the identity on Y , so we can get control in Y . We can also use the fact that these pushes preserve the complement of $Y \times \mathbf{R}$ to obtain control with respect to distance from Y . Together these give (essentially) a controlled h -cobordism $W \rightarrow Y \times (0, \infty)$, with local fundamental group the same as $p \times 1: \text{holink}(X, Y) \times (0, \infty) \rightarrow Y \times (0, \infty)$. The obstruction group for this simplifies somewhat;

$$H_1^{lf}(Y \times (0, \infty); \mathbf{Wh}(p \times 1)) \cong H_0^{lf}(Y; \mathbf{Wh}(p)).$$

2.2 THEOREM. *Suppose $Y \times \mathbf{R} \subseteq Z$ has a manifold factoring of a neighborhood $X \times \mathbf{R} \rightarrow Z$, with $\dim(X - Y) \geq 6$, Y tame and $p: \text{holink}(X, Y) \rightarrow Y$ "reasonable." Then isotopy classes of manifold factorings are classified by the h -cobordism obstruction in $H_0^{lf}(Y; \mathbf{Wh}(p))$.*

Again triviality of the obstruction allows alignment of a countable number of levels, and the pseudoisotopy theorem is used to fill in between these. The method used in this proof appeared in [6, part III, 2.5].

3. Stratified sets. A stratified set is a filtered space $X^n \supseteq X^{n-1} \supseteq \cdots \supseteq X^0$ such that the strata $X^i - X^{i-1}$ are manifolds, and neighborhoods of a stratum in a larger stratum satisfy certain regularity conditions. In the geometric (strongly) stratified sets if $r > i$ the subset $(X^i - X^{i-1}) \subseteq (X^r - X^{r-1}) \cup (X^i - X^{i-1})$ is required to have a mapping cylinder neighborhood with map a fiber bundle, or block bundle. Further there are compatibility conditions on these maps at points where three or more strata come together. All this data gives a detailed geometric picture of neighborhoods in such sets.

A *weakly stratified set* is the homotopy analog of this; $(X^i - X^{i-1}) \subseteq (X^r - X^{r-1}) \cup (X^i - X^{i-1})$ is required to be tame, and have homotopy link a fibration [10]. Conveniently, no compatibility conditions are necessary. This homotopy data does not directly give a geometric picture, but it abundantly provides the sort of hypotheses needed to apply controlled topology. So for example we immediately get obstructions in $H_i^{lf}(X^{n-1}; \mathbf{Wh}(\text{holink}(X^n, X^{n-1})))$ to the existence ($i = 0$) and uniqueness ($i = 1$) of mapping cylinder neighborhoods of X^{n-1} in X^n . A more delicate analysis using the same tools yields a very useful isotopy extension theorem without obstructions [9, 10].

We indicate how to apply the controlled h -cobordism theorem stratum-by-stratum to obtain a theorem for weakly stratified sets. First we permit boundary by defining $\partial X = \bigcup_i \partial(X^i - X^{i-1})$, and require that it be collared in X . Suppose W is a compact weakly stratified set with boundary $\partial W = X_0 \cup X_1$. Then we say W is a weakly stratified h -cobordism if for each j , $(W^{j+1} - W^j)$ deformation retracts to both $(X_0^j - X_0^{j-1})$ and $(X_1^j - X_1^{j-1})$.

Now suppose the stratum W^j is a product $X_0^{j-1} \times I$. We want a product structure on $(W^{j+1} - W^j)$ which fits together with this, so near the lower stratum we need control at least in X_0^{j-1} . More than this, we need control measured in $X_0^{j-1} \times (0, 1]$, with the second coordinate essentially the distance from the lower stratum. The reason for this is we need better and better control measured in X_0^{j-1} the closer we get to it, and for this we use a control function $\delta: X_0^{j-1} \times (0, 1] \rightarrow (0, \infty)$ which goes to 0 as the second coordinate goes to 0. In the interior of the stratum no control is necessary, so we use a point as control space. Putting these together we see the appropriate total control space is the open cone $X_0^{j-1} \times (0, 1]/(X_0^{j-1} \times \{1\})$. (More precisely we need the union of open cones on the components of X_0^{j-1} .) The appropriate map to use to record local fundamental groups is the pushout

$$\begin{array}{ccccc} \text{holink}(X^j, X^{j-1}) & \longrightarrow & X^j - X^{j-1} & \longrightarrow & E \\ \downarrow p & & \downarrow & & \downarrow q \\ X^{j-1} & \longrightarrow & \text{pt} & \longrightarrow & \text{cone}(X^{j-1}) \end{array}$$

3.1 THEOREM [10]. *Suppose W is a weakly stratified h -cobordism of X_0 as above. Then there is an invariant $\tau(W, X_0) \in \sum_i H_1^{lf}(\text{opencone}(X_0^{i-1}); \mathbf{Wh}(q))$. A product structure on the 5-skeleton extends to a product structure on W if and only if the invariant vanishes.*

The argument above indicates how the i th component of the invariant is defined once there is a product structure on the $i - 1$ skeleton. Part of the assertion is that the invariants are well defined without this hypothesis.

Currently the major application is to quotients M/G , to obtain an equivariant topological h -cobordism theorem. Many quotients qualify; if G is finite, each fixed set M^H is a manifold, and none is codimension 2 in another, then M/G is weakly stratified [10].

In the group action case we can give a more universal description of the local fundamental group systems in the obstruction group. Denote the Borel construction by $q_G: (M \times E_G)/G \rightarrow M/G$, and let M^* denote the points with nontrivial isotropy subgroups (so M^*/G is the $(n - 1)$ -skeleton of M/G). There is a fiber map from $p: \text{holink}(M/G, M^*/G) \rightarrow M^*/G$ to the restriction of q_G over M^*/G , which is $(\delta, 1)$ -connected for all $\delta > 0$ if M^* has codimension at least 3. Such an approximately 1-connected fiber map induces an isomorphism of coefficient spectra $\mathbf{Wh}(p) \rightarrow \mathbf{Wh}(q_G)$, so we can use q_G in the top dimensional obstruction group. A convenient way to think of this is that over a point x we get the isotropy group G_x . Generally for a point x , the local fundamental group of the complement of the fixed set M^{G_x}/G , in M^H/G , is $(NH \cap G_x)/H$, where NH denotes the normalizer in G .

4. Rigidity. Certain geometric objects, for example discrete subgroups of some Lie groups, or equivalently their quotient symmetric spaces, have been found to have remarkable uniqueness properties. The rigidity theorem of Mostow is the prototypical example of this. It is becoming clear, however, that the geometric rigidity is only the tip of the phenomenon; it extends deeply into topology and algebra as well. We begin with a topological version.

4.1 CONJECTURE (TOPOLOGICAL RIGIDITY). *Suppose M is a complete Riemannian manifold with nonpositive curvature, G is a discrete group of isometries with finite volume quotient, proper, and codimensions of fixed sets in each other at least 3. Then any proper stratum-preserving homotopy equivalence $X \rightarrow M/G$, X a weakly stratified set, is proper stratum-preserving homotopic to a homeomorphism.*

We hasten to point out that this is certain to be false in some cases when G has torsion; nontrivial examples of Nil or UNil groups should yield counterexamples. The point is that most of the obstructions to homeomorphism encountered in a naive analysis of the problem should vanish or cancel, so the conjecture should be much closer to being true than one would at first expect. We formulate algebraic conjectures, and explain how they are related to the topological version.

For simplicity suppose M/G is compact. As above let $q_G: (M \times E_G)/G \rightarrow M/G$ be the projection in the Borel construction. The diagram

$$\begin{array}{ccc} (M \times E_G)/G & \longrightarrow & (E_G)/G \\ \downarrow q_G & & \downarrow \\ M/G & \longrightarrow & \text{pt} \end{array}$$

induces a homomorphism in homology $H_i(M/G; \mathbf{A}(q_G)) \rightarrow H_i(\text{pt}; \mathbf{A}((E_G)/G)) \cong A_i(G)$, where \mathbf{A} is a spectrum-valued functor (**Wh** above).

4.2 CONJECTURE (ALGEBRAIC RIGIDITY). *Suppose M, G are as above, M/G is compact, and \mathbf{A} denotes either the algebraic K -theory spectrum, or an appropriate L -theory spectrum. Then the natural homomorphism*

$$H_i(M/G; \mathbf{A}(r)) \rightarrow A_i(G)$$

is an isomorphism.

Again this is false in some cases where G has torsion, but it is torsion which makes it most interesting; the torsion appears as isotropy subgroups, therefore as the fundamental groups in inverse images of r . The suggestion is that the K or L theory of G , an infinite group, is determined in terms of the K or L theory of the finite subgroups.

The algebraic conjecture for L theory implies the geometric conjecture. We illustrate the method by showing the somewhat easier fact that the algebraic conjecture for K -theory implies: any weakly stratified h -cobordism of M/G is topologically a product. Denote M/G by X , to match notation with §3. Then Theorem 3.1 shows that this assertion follows from the vanishing for all i of the groups $H_1^{lf}(\text{opencone}(X^{i-1}); \mathbf{Wh}(q))$. The locally finite homology is isomorphic to the relative homology $H_1(\text{cone}(X_0^{i-1}), X_0^{i-1}; \mathbf{Wh}(q))$, so the cofibration diagram used to define the map q gives a long exact sequence

$$\begin{aligned} \cdots \rightarrow H_j(X^{i-1}; \mathbf{Wh}(p)) &\rightarrow H_j(\text{pt}; \mathbf{Wh}(\pi_1 X^i - X^{i-1})) \\ &\rightarrow H_j(\text{cone}(X^{i-1}), X^{i-1}; \mathbf{Wh}(q)) \rightarrow \cdots \end{aligned}$$

The vanishing of the relative group (for all j) is equivalent to the first homomorphism being an isomorphism. We claim that this homomorphism is one of those specified in the conjecture, with coefficient **Wh** rather than **K**. (The Whitehead spectrum is derived from the K spectrum, and the conjectures for the two spectra are equivalent.) For simplicity we concentrate on the top stratum, $X^i = M/G$ and $X^{i-1} = M^*/G$.

The middle group is the homotopy group of the spectrum,

$$\text{Wh}_i(\pi_1(M/G - M^*/G)).$$

The codimension 3 hypothesis implies $M - M^*$ is simply connected, so the quotient has fundamental group G , and the group is $\text{Wh}_i(G)$. In the first group, according to the analysis at the end of §3, we can replace the homotopy link map in the coefficient system by the projection in the Borel construction. Therefore it is $H_j(M^*/G; \mathbf{Wh}(q_G))$. This is isomorphic to the homology of all

of $M/G, H_j(M/G; \text{Wh}(q_G))$, because over $(M - M^*)/G$ the coefficients vanish (the fiber of r is 1-connected there, and $\text{Wh}_i(\{1\}) = 0$). The groups are therefore the ones in the conjecture, and it is easily seen that the homomorphism is the same as well.

As remarked above, the conjecture as formulated in 4.2 is false. The next statement gives versions we can hope to be true.

4.3 CONJECTURE. *Suppose M, G as in Conjecture 4.2, and $\mathbf{Q} \supseteq R$ is a subring.*

(1) $H_i(M/G; \mathbf{K}(r; R)) \rightarrow K_i(R[G])$ is an isomorphism if orders of torsion elements in G are invertible in R , and for any R is an isomorphism modulo order of torsion.

(2) $H_i(M/G; \mathbf{L}(r; R)) \rightarrow L_i(R[G])$ is an isomorphism if orders of torsion elements in G are invertible in R , and for any R is an isomorphism modulo 2-torsion.

A tremendous amount of work has been done on special cases of these, particularly on (2) in the weak form that the homomorphism is a rational injection (also known as the "Novikov Conjecture"). We cannot detail all this work, but will mention some of the current high points. In K -theory the first substantial progress was due to Farrell and Hsiang [2], who proved Conjecture 4.3 for almost flat M , torsion-free G , and $i = 0, 1$. This was extended to all i and allowing torsion, when M is a certain type of symmetric space and orders of torsion are invertible in R , in [7]. It has very recently been proved in the torsion-free hyperbolic case by Farrell and Jones [5]. In L -theory Farrell and Hsiang have proved Conjecture 4.3(2) in the almost flat torsion-free case [3], and Yamasaki [11, 12] has extended this mod 2-torsion to allow torsion in G . Farrell and Hsiang have also proved that the homomorphism is a split injection, in many torsion-free cases [4].

All of these partial results involve controlled topology in their proofs. Various algebraic and geometric manipulations yield ε control, and then the characterization theorem mentioned at the end of §1 is used to identify the controlled group as homology.

It seems likely, particularly after the dramatic results announced by Farrell and Jones, that Conjecture 4.3 will be proved before long. Understanding the discrepancy between Conjectures 4.3 and 4.2 (Nil and UNil groups), which will be necessary for a full analysis of the geometric situation of Conjecture 4.1, seems further off.

BIBLIOGRAPHY

1. T. A. Chapman, *Controlled simple homotopy theory and applications*, Lecture Notes in Math., vol. 1009, Springer-Verlag, 1983.
2. F. T. Farrell and W.-C. Hsiang, *The Whitehead group of poly-(finite or cyclic) groups*, J. London Math. Soc. (2) **14** (1981), 308–324.
3. ———, *Topological characterization of flat and almost flat Riemannian manifolds M^n ($n \neq 3, 4$)*, Amer. J. Math. **105** (1983), 641–672.

4. —, *On Novikov's conjecture for cocompact discrete subgroups of a Lie group*, Lecture Notes in Math., vol. 1051, Springer-Verlag, 1984, pp. 38–48.
5. F. T. Farrell and L. E. Jones, *Algebraic K-theory of hyperbolic manifolds*, Bull. Amer. Math. Soc. (N.S.) **14** (1986), 115–120.
6. F. Quinn, *Ends of maps*. I, Ann. of Math. (2) **110** (1979), 275–331; II, Invent. Math. **68** (1982), 353–424; III, J. Differential Geom. **17** (1982), 503–521; IV, Amer. J. Math. **108** (1986), 1139–1162.
7. —, *Algebraic K-theory of poly-(finite or cyclic) groups*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), 221–226.
8. —, *Geometric algebra*, Lecture Notes in Math., vol. 1126, Springer-Verlag, 1985, pp. 182–198.
9. —, *Intrinsic skeleta and intersection homology of weakly stratified sets*, Geometry and Topology: Manifolds, Varieties, McCrory and Shifrin, editors, Marcel Dekker, New York, 1986, pp. 233–249.
10. —, *Weakly stratified sets*, Preprint, 1985.
11. M. Yamasaki, *Surgery groups of crystallographic groups*, Thesis, Virginia Polytechnic Inst. and State Univ., Blacksburg, Va., 1982.
12. —, *L-groups of virtually polycyclic groups*, Preprint, 1986.

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY, BLACKSBURG, VIRGINIA 24061, USA

Representations of 3-Manifold Groups and Applications in Topology

PETER B. SHALEN

I am going to describe a technique in 3-dimensional topology which first appeared in my joint work with M. Culler [2, 3], and which was further exploited in my joint work with J. Morgan [7, 8] and with Culler, C. McA. Gordon, and J. Luecke [1]. The technique has a broad spectrum of applications in 3-manifold theory, in the study of knots, of group actions, of surfaces in 3-manifolds, and of hyperbolic metrics. It is based on an interaction between two of the most important themes in 3-manifold theory: incompressible surfaces, as used in the work of Stallings, Haken, Waldhausen, and others, and hyperbolic geometry, as used most strikingly in the work of Thurston. As I shall explain below, both 3-dimensional hyperbolic geometry and incompressible surface theory are related to the study of certain kinds of “representations” of the fundamental group of a 3-manifold. The link between the two kinds of representations is provided by the Bruhat-Tits building for SL_2 .

1. Hyperbolic manifolds. A *hyperbolic manifold* is a complete, connected Riemannian manifold of constant sectional curvature -1 . The only simply connected hyperbolic n -manifold is *hyperbolic n -space* H^n , the non-Euclidean space of Bolyai, Lobachevski, and Gauss. Thus if M is an oriented hyperbolic n -manifold, the universal cover of M may be identified with H^n , and $\pi_1(M)$ with a discrete subgroup of the group $SO(n, 1)$ of orientation-preserving isometries of H^n . The “discrete,” faithful representation of $\pi_1(M)$ in $SO(n, 1)$ constructed in this way is well defined up to equivalence.

The group $SO(3, 1)$ may be identified with $PSL_2(\mathbb{C})/\{\pm I\}$, and for a hyperbolic 3-manifold M , the natural representation of $\pi_1(M)$ in $PSL_2(\mathbb{C})$ may be shown to lift (though not quite uniquely) to a representation in $SL_2(\mathbb{C})$. Thus for a connected, orientable 3-manifold M , the study of representations of $\pi_1(M)$ in $SL_2(\mathbb{C})$ is closely related to the study of hyperbolic metrics on M .

If M is an oriented hyperbolic 3-manifold whose *volume* (as a Riemannian manifold) is finite, then certain topological properties of M are easily established. First, M is diffeomorphic to the interior of a compact manifold N whose

boundary components are tori. Second, N is *irreducible* in the sense that every smooth 2-sphere in N bounds a 3-ball. Third, for each component B_i of ∂N , the inclusion homomorphism maps $\pi_1(B_i)$ isomorphically onto a subgroup H_i of $\pi_1(N)$, and each rank-two free abelian subgroup of $\pi_1(N)$ is contained in a conjugate of one of the H_i . A very deep result of Thurston's implies that if M is noncompact and has the three properties stated above, and M is not diffeomorphic to $T^2 \times \mathbf{R}$ or to a twisted \mathbf{R} -bundle over a Klein bottle, then M admits a hyperbolic metric of finite volume (which, by Mostow rigidity [9], is then essentially unique).

Many noncompact, connected, oriented 3-manifolds arising naturally in applications have the first two of the three properties stated above: the simplest example is the complement of an arbitrary knot in S^3 . By combining Thurston's theorem with the decomposition theorem of Johannson [5] and Jaco-Shalen [4], one can essentially reduce many questions about such manifolds to the case of hyperbolic manifolds of finite volume.

2. Incompressible surfaces. Let N be a compact, connected, orientable 3-manifold (possibly with boundary). Consider a connected, compact, orientable surface $S \subset N$ which is *properly embedded* in the sense that $S \cap \partial N = \partial S$. One says that S is *incompressible* in M if the inclusion homomorphism $\pi_1(S) \rightarrow \pi_1(N)$ is injective and if, in addition, S is nondegenerate in a suitable sense. For my purposes, the nondegeneracy conditions will be that S is not a 2-sphere, and that S is not *boundary-parallel*, i.e., there is no set $P \subset N$ with frontier S such that the pair (P, S) is homeomorphic to $(S \times [0, 1], S \times \{0\})$. I will say that a possibly disconnected compact surface $S \subset M$ is *incompressible* if its components are orientable and are all incompressible in the above sense.

Given a nonempty, properly embedded, orientable surface $S \subset M$, we can consider the universal cover $p: \tilde{M} \rightarrow M$ and the surface $\tilde{S} = p^{-1}(S) \subset \tilde{M}$. There is a 1-dimensional simplicial complex that is "dual" to \tilde{S} : it has a vertex v_c for each component c of $\tilde{M} - \tilde{S}$, an edge e_s for every component s of \tilde{S} , and we have $v_c < e_s$ iff $s \subset \bar{c}$. This 1-complex T is a *tree*, i.e., it is 1-connected. There is a natural simplicial action of $\pi_1(M)$ of T . This action has no *inversions*: that is, if an element of $\pi_1(M)$ stabilizes an edge e then it fixes both vertices incident to e . The assumption $S \neq \emptyset$ implies that no point of T is fixed by (all of) $\pi_1(M)$. I will express this by saying that the action is *nontrivial*.

We may regard an action of $\pi_1(M)$ on the tree defined by the surface S as a kind of nonlinear representation of $\pi_1(M)$ associated with S . If M is irreducible, *incompressible* surfaces are classified by the associated representation; i.e., two incompressible surfaces are isotopic iff there is a $\pi_1(M)$ -equivariant simplicial isomorphism between their associated trees.

This already shows that the study of $\pi_1(M)$ on trees is related to the study of incompressible surfaces in M . Actually, the connection is much closer than the above discussion might suggest. It is not true that every action of $\pi_1(M)$ on a tree can be defined by a surface; nevertheless, for M irreducible, there is

a very useful (if noncanonical) way of constructing incompressible surfaces from general actions of $\pi_1(M)$ on trees. Namely, if $\pi_1(M)$ acts (simplicially, without inversions) on a tree T , there is always a $\pi_1(M)$ -equivariant map $\tilde{f}: \tilde{M} \rightarrow T$, and \tilde{f} may be taken to be transverse to the set \tilde{Q} of midpoints of edges of T . Now \tilde{f} induces a map $f: M \rightarrow T/\pi_1(M)$, transverse to the image Q of \tilde{Q} in T , and $S = f^{-1}(Q)$ is a properly embedded, orientable surface in M . If the action of $\pi_1(M)$ on T is nontrivial, then for any choice of \tilde{f} we have $S \neq \emptyset$. On the other hand, using fundamental results due to Papakyriakopoulos, one can show that \tilde{f} may always be chosen so that S is incompressible in M .

The construction just described is due to Stallings, who exploited it extensively in the early 1960s. (He did not work in terms of trees, but used a different point of view which, via the Bass-Serre theory [10] of groups acting on trees, is equivalent to the point of view taken here.)

3. Trees and the space of characters. Let M be a connected, compact, oriented 3-manifold (possibly with boundary). I have explained that the study of representations of $\pi_1(M)$ in $\mathrm{SL}_2(\mathbb{C})$ is related to the study of hyperbolic metrics on $\mathrm{int} M$, and that the study of actions of $\pi_1(M)$ on trees is related to the study of incompressible surfaces in M . Next I will explain how, for any finitely generated group Γ , representations of Γ in $\mathrm{SL}_2(\mathbb{C})$ are related to actions of Γ on trees: this provides the basic interaction between incompressible surfaces and hyperbolic geometry.

Let $\gamma_1, \dots, \gamma_n$ be generators for Γ . A representation ρ of Γ in $\mathrm{SL}_2(\mathbb{C})$ is determined by the n -tuple of matrices $(\rho(\gamma_1), \dots, \rho(\gamma_n))$. This n -tuple may be regarded as a point of the $4n$ -dimensional complex affine space M^n , where M is the space of all 2×2 complex matrices. Thus we may identify the set of all representations of Γ in $\mathrm{SL}_2(\mathbb{C})$ with a subset $R(\Gamma)$ of M^n . It is easy to see that $R(\Gamma)$ is a closed complex affine algebraic set in the affine space M^n .

Whereas the points of the affine algebraic set $R(\Gamma)$ are linear representations of Γ , "ideal" points of $R(\Gamma)$ may often be interpreted as corresponding to nonlinear representations, namely actions on trees. This assertion can be made precise in various ways; the most general way is the theory of [7]. What I shall describe here is a simpler way which is presented in [2] and is sufficient for many applications. Instead of considering the entire algebraic set $R(\Gamma)$, we consider a (complex, affine, irreducible) curve $C \subset R(\Gamma)$. If F denotes the function field of C , there is a "universal" representation $P: \Gamma \rightarrow \mathrm{SL}_2(F)$. (Given $\gamma \in \Gamma$, we have $P(\gamma) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, where $a, b, c, d \in F$ are the functions on C determined by $\rho(\gamma) = \begin{pmatrix} a(\rho) & b(\rho) \\ c(\rho) & d(\rho) \end{pmatrix}$ for any point $\rho \in C$; here we regard ρ as a representation of Γ in $\mathrm{SL}_2(\mathbb{C})$.)

On the other hand, let \hat{C} denote a projective completion of the affine curve C in which all ideal points are smooth. The function field of \hat{C} is identified with F . Hence a smooth point $x \in \hat{C}$ determines a (discrete, rank-1) valuation $v: F^* \rightarrow \mathbb{Z}$ of the field F . (For $f \in F^*$ we have $v(f) = n$ if f has a zero of order n at x , $v(f) = -n$ if f has a pole of order n , and $v(f) = 0$ if f has neither a pole nor a zero at x .)

In general, if F is a field with a valuation v , and P is a representation of a group Γ in $\mathrm{SL}_2(F)$, there is a standard way of constructing an action of Γ on a tree. This tree T is an example of a Bruhat-Tits building; the following description of T is due to Serre [10]. Let $\mathcal{O} \subset F$ denote the valuation ring consisting of all $f \in F$ such that $f = 0$ or $v(f) \geq 0$. The unique maximal ideal \mathcal{M} of \mathcal{O} consists of all $f \in F$ such that $f = 0$ or $v(f) > 0$. A *lattice* in the vector space F^2 is a finitely generated \mathcal{O} -module spanning F^2 . Two lattices L, L' are *equivalent* if $L' = \alpha L$ for some $\alpha \in F^*$. A *vertex* of T is an equivalence class of lattices. Two vertices are joined by an *edge* if and only if they can be represented by lattices L, L' with $L' \subset L$, and L/L' isomorphic as an \mathcal{O} -module for \mathcal{O}/\mathcal{M} . A representation $P: \Gamma \rightarrow \mathrm{SL}_2(F)$ is plainly seen to induce a simplicial action of Γ on the 1-complex T . In [10] it is shown that T is 1-connected and that Γ acts without inversions.

Thus every smooth point of \hat{C} gives rise to an action of Γ on a tree. However, it is easy to see that smooth points of $C \subset \hat{C}$ give rise to trivial actions in the sense of §2. Even for *ideal* points of \hat{C} , it is not obvious whether the corresponding actions are trivial.

The picture becomes clear if we modify our point of view slightly. Each point of $R(\Gamma)$ is a specific representation of Γ in $\mathrm{SL}_2(\mathbb{C})$; however, it is natural to regard two representations as being “the same” if they are equivalent. The groups $\mathrm{GL}_2(\mathbb{C})$ act by conjugation on $R(\Gamma)$, and the orbits are equivalence classes of representations. In the category of complex algebraic sets there is a well-defined quotient $X(\Gamma)$ of $R(\Gamma)$ by the action of $\mathrm{GL}_2(\mathbb{C})$. Moreover, $X(\Gamma)$ is a closed affine algebraic set, the natural map $t: R(\Gamma) \rightarrow X(\Gamma)$ is surjective and maps each orbit to a point, and for an irreducible representation ρ , $t^{-1}(t(\rho))$ is precisely the orbit of ρ . In general, given $\rho, \rho' \in R(\Gamma)$ we have $t(\rho) = t(\rho')$ if and only if $\mathrm{trace} \rho(\gamma) = \mathrm{trace} \rho'(\gamma)$ for every $\gamma \in \Gamma$. (Thus $X(\Gamma)$ is a “space of characters” for representations of Γ in $\mathrm{SL}_2(\mathbb{C})$.) In particular, there is a well-defined function $I_\gamma: X(\Gamma) \rightarrow \mathbb{C}$ given by $I_\gamma(t(\rho)) = \mathrm{trace} \rho(\gamma)$. The functions I_γ , $\gamma \in \Gamma$, belong to the coordinate ring of $X(\Gamma)$ and generate it as a \mathbb{C} -algebra.

Now let $C \subset X(\Gamma)$ be a curve and \hat{C} a projective completion of C in which all ideal points are smooth. A slight modification of the above construction allows one to associate with any ideal point of \hat{C} a nontrivial action of Γ on a tree. This is the basic connection between representation in $\mathrm{SL}_2(\mathbb{C})$ and actions on trees which was first formulated and applied to 3-manifold theory in [2]. I shall explain some of the applications below.

4. From hyperbolic metrics to incompressible surfaces. Let M be a hyperbolic 3-manifold of finite volume, and let N denote the compactification described in §1. The hyperbolic metric on N gives rise to a discrete, faithful representation of $\pi_1(M) = \pi_1(N)$ in $\mathrm{PSL}_2(\mathbb{C})$ which, as I mentioned in §1, lifts to a representation ρ_0 in $\mathrm{SL}_2(\mathbb{C})$; thus ρ_0 is a point of $R(\pi_1(M))$, and $x_0 = t(\rho_0)$ is a point of $X(\pi_1(M))$. It was first shown by Thurston that x_0 is a smooth

point of a component X_0 of $X(\pi_1(M))$ and that the (complex) dimension of X_0 is equal to the number of (torus) components of ∂N .

Thus if M is noncompact we have $\dim X_0 > 0$, and hence X_0 contains at least one curve. For any curve C in X_0 , if \hat{C} is the canonical completion described in §3, the ideal points of \hat{C} give rise to action of $\pi_1(M)$ on trees. These in turn may be used to construct incompressible surfaces as in §2.

Thus we have a natural method for constructing incompressible surfaces in hyperbolic 3-manifolds. Using Thurston's theorem as indicated in §1, one can extend this to a method for constructing incompressible surfaces in any compact, connected, irreducible, orientable 3-manifold whose boundary is nonempty and consists of tori. By this method one can show, for example, that if K is a nontrivial knot in S^3 , then the complement N of a tubular neighborhood of K in S^3 contains an incompressible surface with nonempty boundary which divides N into two components. This result, which was proved in my joint paper with Culler [3], established an old conjecture of Neuwirth's on the structure of knot groups: every knot group is a nontrivial free product with a free amalgamated subgroup.

5. Group actions. The ideas described above were applied in [2] to the study of finite group actions on 3-manifolds. To illustrate these applications, I will sketch a proof of the so-called "Smith Conjecture": every periodic diffeomorphism h of S^3 which preserves orientation and has nonempty fixed point set is conjugate (in $\text{Diff}(S^3)$) to a rotation (in $\text{SO}(4)$). (For a history of this result and an account of the first proof, see [6].)

We may assume that the period n of h is > 1 . If Σ denotes the quotient of S^3 by the action of the cyclic group generated by h , then it follows from old work of P. A. Smith's that Σ is a closed orientable 3-manifold, that the projection $p: S^3 \rightarrow \Sigma$ is a branched covering map, and that the branch set $K \subset \Sigma$ is diffeomorphic to S^1 . The Smith Conjecture may be seen to be equivalent to the assertion that $M = \Sigma - K$ is diffeomorphic to $S^1 \times \mathbf{R}$. We assume that this is false and derive a contradiction. Using Thurston's theorem as indicated in §1, one can show that the crucial case is the one in which M has a hyperbolic metric of finite volume. It is this case that I shall discuss. For simplicity I will also assume $n > 2$.

Let H denote a tubular neighborhood of K in Σ . Then $N = \overline{\Sigma - H}$ may be identified with the compactification of M discussed in §1. Let $\mu \subset \partial N$ be a simple closed curve that bounds a disk in H meeting K transversely in one point. Up to conjugacy, μ defines an element of $\pi_1(M)$ which I will also call μ . Using covering space theory and the simple connectivity of S^3 one sees that $[\pi_1(M): \mu^n = 1]$ is a cyclic group of order n .

Now consider the component X_0 of $X(\pi_1(M))$ defined in §4; in this case X_0 is a curve. Recall from §3 that there is a function $I_\mu: X_0 \rightarrow \mathbf{C}$ defined by $I_\mu(t(\rho)) = \text{trace } \rho(\mu)$ for $t(\rho) \in X_0$. If \hat{X}_0 is the projective completion of X_0 whose ideal points are all smooth, then I_μ extends to a function $\hat{I}_\mu: \hat{X}_0 \rightarrow \mathbf{C} \cup \{\infty\}$. Note

that either (a) I_μ takes every complex value on X_0 , or (b) $\hat{I}_\mu(x_0) \neq \infty$ for some ideal point x_0 of \hat{X}_0 . In case (a) one can show (using $n > 2$) that some point of X_0 has the form $t(\rho)$ where $\rho(\pi_1(M))$ is cyclic and $\rho(\mu)$ has order n . This contradicts the fact that $|\pi_1(M): \mu^n = 1|$ is cyclic. In case (b) we can construct, as in §4, an incompressible surface $S \subset N$ from the ideal point x , and it may be shown (using $\hat{I}_\mu(x) \neq \infty$) that the boundary components of S (if any) are homotopic to μ in ∂N . Hence we may cap the components of ∂S with disks to form a closed surface $F \subset \Sigma$. An argument due to Gordon and Litherland [6], based on work of Meeks and Yau [6] allows one to replace F by a surface F' such that $p^{-1}(F')$ is either an incompressible surface or a nonseparating 2-sphere; in either case, the simple connectivity of S^3 is contradicted.

6. Knots and surgery. It is a classical conjecture that knots in S^3 are determined up to unoriented equivalence by the topological type of their complements (or, equivalently, of the complements of their tubular neighborhoods). In my joint paper with Culler, Gordon, and Luecke [1], it is shown that up to unoriented equivalence there are at most two knots whose complements are of a given topological type. Moreover, many knots are shown to be uniquely characterized by their complements. These and a number of related results about Dehn surgery on knots, as well as W. Whitten's result [11] that prime knot complements are classified by their fundamental groups, are applications of the main theorem of [1]. To understand the setting of this theorem, notice that if N is the complement of an open tubular neighborhood of a knot in S^3 , we can describe S^3 as the union of N with a solid torus $D^2 \times S^1$, glued along their boundaries by a homeomorphism. The type of the knot is easily seen to be determined by the *meridian* of the knot, i.e., the simple loop $\gamma \subset \partial N$ which is identified with $\partial D^2 \times \{\text{point}\}$ under the gluing homeomorphism.

Now let N be a compact, orientable, irreducible 3-manifold whose boundary is a torus. For any nontrivial simple loop $\gamma \subset \partial M$ we may construct a closed manifold N_γ by attaching a solid torus $D^2 \times S^1$ to N by a homeomorphism between their boundaries that maps γ onto $(\partial D^2) \times \{\text{point}\}$. The (unoriented) isotopy class of γ determines N_γ up to homeomorphism.

Suppose that N is not a union of two solid tori meeting along an annulus. The theorem then asserts that for any two nontrivial simple loops γ, γ' such that $\pi_1(N_\gamma), \pi_1(N_{\gamma'})$ are cyclic, the homological intersection number of γ and γ' in ∂N is at most 1. It follows easily that there are at most three simple loops γ , up to isotopy, for which $\pi_1(N_\gamma)$ is cyclic. Fintushel-Stern and Berge have produced infinitely many examples in which three such loops do exist. The theorem also easily implies that there are at most two simple loops γ for which $\pi_1(N_\gamma)$ is trivial. This gives the result that there are at most two knots with a given complement.

I want to sketch the proof of the theorem, emphasizing the role of the methods I have discussed above. For simplicity I will discuss only the case where N contains no closed incompressible surfaces. In particular, N contains no

incompressible torus; and one can then show, using Thurston's theorem, that $M = \text{int } N$ has a hyperbolic metric of finite volume. Thus, as in §4 we may define a component X_0 of $X(\pi_1(M))$; as in §5 we have $\dim X_0 = 1$, and we may again consider a projective completion \hat{X}_0 of X_0 in which all ideal points are smooth.

Consider the group $L = H_1(\partial N; \mathbf{Z})$. We may identify L with $\pi_1(\partial N)$ and hence with a subgroup of $\pi_1(N)'$. Thus for each $\alpha \in L$ there is a function $I_\alpha: X_0 \rightarrow \mathbf{C}$, which extends to a function $\hat{I}_\alpha: \hat{X}_0 \rightarrow \mathbf{C} \cup \{\infty\}$. The degree of I_α , $\deg I_\alpha = \deg \hat{I}_\alpha$, is a well-defined integer depending on α . It is shown in [1] that there is a norm $\| \cdot \|$ on the vector space $V = H_1(\partial N; \mathbf{R})$ such that for every element α of the lattice $L \subset V$ we have $\|\alpha\| = \deg I_\alpha$. The unit ball in this norm is a polygon.

Set $m = \min_{0 \neq \alpha \in L} \|\alpha\|$. The key step is to show that if γ is a simple loop in ∂N , which we regard as an indivisible element of L defined up to sign, we have $\|\gamma\| = m$. In sketching the proof of this, I will simplify the language by assuming that X_0 is smooth. If $\|\gamma\| > m$ then $\deg f_\gamma > \deg f_\delta$ for some $\delta \in L - \{0\}$, where $f_\gamma = I_\gamma^2 - 4$. Hence for some $x \in \hat{X}_0$ we have $f_\gamma(x) = 0$, and either $f_\delta(x) \neq 0$ or f_δ has a zero of order smaller than the zero of f_γ . If $x \in X_0$ then it can be shown that $x = t(\rho)$ for some $\rho \in R(\pi_1(M))$ with $\rho(\gamma) = \pm I$; hence ρ induces a representation $\bar{\rho}$ of $\pi_1(N_\gamma) = |\pi_1(N): \gamma = 1|$ in $\text{PSL}_2(\mathbf{C})$. Furthermore, ρ may be chosen so that $\bar{\rho}(\pi_1(N_\gamma))$ is noncyclic, which is absurd since $\pi_1(N_\gamma)$ is cyclic. If x is an ideal point of \hat{X}_0 then as in §4 we may use x to construct an incompressible surface $S \subset N$; the properties of x imply that the boundary components of S are homotopic to γ in ∂M . Thus we may cap the components of ∂S with disks to obtain a closed surface $\hat{S} \subset \partial M_\gamma$. Under our simplifying assumption that N contains no closed incompressible surfaces, it is then possible to replace \hat{S} by a closed surface \hat{S}' such that either \hat{S}' is incompressible in N_γ or \hat{S}' is a sphere that decomposes N_γ as a connected sum of two nontrivial lens spaces. In either case we again contradict the cyclicity of $\pi_1(N_\gamma)$.

Thus any such γ lies on the boundary of the ball B of radius m in the normed vector space V . A refinement of the above argument shows that γ is not a vertex of the polygon B . Since the interior of B contains no points of $L - \{0\}$, a lemma of Minkowski's shows that the area of B (in the natural measure on $V = H_1(\partial N; \mathbf{R})$) is at most 4. Now if $\gamma, \gamma' \in L$ both lie on the boundary of B then $\pm\gamma$ and $\pm\gamma'$ span a parallelogram $Q \subset L$; if γ, γ' are not vertices of B then the area of Q is strictly less than 4. But the numerical intersection number of γ and γ' is one-half the area of Q and is therefore at most 1, as the theorem asserts.

7. Variation of hyperbolic structure. In the applications discussed above, the existence of a finite-volume hyperbolic metric on a manifold M was used to produce an appropriate curve in $X(\pi_1(M))$ to get the argument started. In my joint papers with Morgan [7, 8], the method is applied in a quite different way: one begins with a hyperbolic manifold of infinite volume and studies the subset

\mathcal{D} of $X(\pi_1(M))$ consisting of points defined by discrete, faithful representations of $\pi_1(M)$ in $\mathrm{SL}_2(\mathbb{C})$. These are very closely related to different hyperbolic structures on M . (In the finite-volume case, classical Weil rigidity implies that \mathcal{D} is merely a finite set.) A deep result of Thurston's gives necessary and sufficient conditions for \mathcal{D} to be compact. It is natural to try to prove this theorem by the methods described in §§2 and 3 above, since a noncompact set "goes out to infinity" and—according to §§2 and 3—the "structure at infinity" of $X(\pi_1(M))$ should be related to actions of $\pi_1(M)$ on trees and thus to incompressible surfaces. Morgan and I succeeded in carrying out this program in [7] and [8].

However, since $X(\pi_1(M))$ is not a curve, we needed to generalize the theory rather elaborately, using nondiscrete valuations and extending the notions of tree and of incompressible surface.

Our immediate reason for wanting to understand Thurston's compactness theorem better was that this theorem is a key step in Thurston's proof of his existence theorem for finite-volume hyperbolic metrics discussed in §1 above. Thus the same methods that I have been describing for exploiting Thurston's existence theorem topologically can be used to understand an important part of the proof of this same existence theorem.

An exposition of the theory that I have alluded to here, and of Morgan's generalization of it to higher-dimensional hyperbolic manifolds, may be found in "Trees and Hyperbolic Geometry," Morgan's contribution to these Proceedings.

REFERENCES

1. M. Culler, C. McA. Gordon, J. Luecke, and P. B. Shalen, *Dehn surgery on knots*, Ann. of Math. (to appear).
2. M. Culler and P. B. Shalen, *Varieties of groups representations and splittings of 3-manifolds*, Ann. of Math. **117** (1983), 109–146.
3. —, *Bounded separating, incompressible surfaces in knot manifolds*, Invent. Math. **75** (1984), 537–545.
4. W. Jaco and P. B. Shalen, *Seifert fibered spaces in 3-manifolds*, Mem. Amer. Math. Soc. **21** (1979), no. 220.
5. K. Johansson, *Homotopy equivalence of 3-manifolds with boundaries*, Lecture Notes in Math., vol. 761, Springer-Verlag, Berlin-New York, 1979.
6. J. W. Morgan and H. Bass (editors), *The Smith conjecture*, Academic Press, New York, 1984.
7. J. W. Morgan and P. B. Shalen, *Valuations, trees, and degenerations of hyperbolic structures. I*, Ann. of Math. **120** (1984), 401–476.
8. —, *ibid.*, part II, Preprint.
9. G. D. Mostow, *Quasi-conformal mappings in n -space and the rigidity of hyperbolic space forms*, Inst. Hautes Études Sci. Publ. Math. **34** (1968), 53–104.
10. J. P. Serre (with H. Bass), *Trees*, Springer-Verlag, 1980.
11. W. Whitten, *Knot complements and groups* (to appear).

UNIVERSITY OF ILLINOIS AT CHICAGO, CHICAGO, ILLINOIS 60680, USA

Geometric Representation Theory of Compact Lie Groups

TAMMO TOM DIECK

A *homotopy representation* (HR) of the compact Lie group G is a G -space Y which is G -homotopy-equivalent to a G -complex [14, Chapter II] X such that: (i) X has finite orbit type; (ii) for each (closed) subgroup $H \subset G$ the H -fixed point set X^H is homotopy-equivalent to a sphere $S^{n(H)}$; (iii) X^H is an $n(H)$ -dimensional space. The HR is called *finite* if X can be chosen as a finite G -complex. (X^H empty $\Leftrightarrow n(H) = -1$). A (spherical) *representation form* (RF) of G is an HR Y for G such that Y^H is isomorphic to $S^{n(H)}$ (in the categories DIFF, PL, or TOP, whenever this makes sense). The unit sphere SV in an orthogonal representation space V is an HR and an RF, called *linear*.

The join $X * Y$ of HRs X and Y is an HR. The assignment $(X, Y) \mapsto X * Y$ induces a composition law on the set of G -homotopy classes of HRs. Let $V(G)$ denote the associated Grothendieck group, written additively.

Let $\psi(G)$ be the space of conjugacy classes (H) of closed subgroups H of G . The assignment $(H) \mapsto n(H) + 1$ is a continuous function $\psi(G) \rightarrow \mathbf{Z}$, called the *dimension function* $\text{Dim } X$ of the HR X . Let $C'(G) = C(\psi(G), \mathbf{Z})$ be the ring of continuous functions $\psi(G) \rightarrow \mathbf{Z}$. We have $\text{Dim}(X * Y) = \text{Dim } X + \text{Dim } Y$ and therefore, by universality of $V(G)$, an additive homomorphism $\text{Dim}: V(G) \rightarrow C'(G)$; its kernel is denoted by $v(G)$.

Let X and Y be HRs with $\text{Dim } X = \text{Dim } Y$ and let $f: X \rightarrow Y$ be a G -map. We choose orientations for all X^H and Y^H . Then the map $f^H: X^H \rightarrow Y^H$ has a degree $d(f^H) \in \mathbf{Z}$. The function $d(f): (H) \mapsto d(f^H)$ is contained in $C'(G)$ and is called the *degree function* of f . Let $\varphi(G) \subset \psi(G)$ be the subspace of classes (H) with finite $WH := NH/H$. It suffices to consider the restriction $d(f) \in C(G) := C(\varphi(G), \mathbf{Z})$.

The *Burnside ring* $A(G)$ of G is defined to be the set of equivalence classes of finite G -complexes with addition (multiplication) induced by disjoint union (cartesian product); the equivalence relation is: $X \sim Y \Leftrightarrow$ for all $H \subset G$, $\chi(X^H) = \chi(Y^H)$ (where χ is the Euler characteristic); see [6]. One can also use finitely dominated G -complexes. The map $\varphi: A(G) \rightarrow C(G), X \mapsto ((H) \mapsto \chi(X^H))$ is

an injective ring homomorphism, treated as identification. It turns $C(G)$ into an $A(G)$ -module. An $A(G)$ -submodule $M \subset C(G)$ is called *invertible* if there exists a submodule N such that $MN = \{\sum m_i n_i | m_i \in M, n_i \in N\}$ equals $A(G)$. Let $\text{Inv } A(G)$ be the abelian group of invertible submodules with composition law $(M, N) \mapsto MN$. An $A(G)$ -module M is called *projective of rank one* if it is finitely generated and if there exists N such that $M \otimes_{A(G)} N \cong A(G)$. Let $\text{Pic } A(G)$, the Picard group of $A(G)$, be the abelian group of projective $A(G)$ -modules of rank one with composition law $(M, N) \mapsto M \otimes_{A(G)} N$. An invertible module is projective of rank one. This fact leads to an exact sequence

$$1 \rightarrow A(G)^* \rightarrow C(G)^* \rightarrow \text{Inv } A(G) \rightarrow \text{Pic } A(G) \rightarrow 1$$

(S^* denotes the unit group of the ring S). Let $o(G)$ be the smallest common multiple of all $|WH|$, $(H) \in \varphi(G)$; see [7] for its existence. Let $\omega(X, Y) = \{d(f) | f: X * SV \rightarrow Y * SV, \text{ some } V\} \subset C(G)$.

PROPOSITION 1. (i) $\omega(X, X)$ is independent of X and equal to the Burnside ring.

(ii) $\omega(X, Y)$ is an invertible $A(G)$ -submodule of $C(G)$ which depends only on the class of $[X] - [Y] \in v(G)$.

(iii) $o(G)\omega(X, Y) \subset A(G)$.

(iv) $\omega(X, Y)$ contains $z \in C(G)$ with values prime to $o(G)$.

THEOREM 2 ([18] FOR FINITE G). The assignment $([X] - [Y]) \mapsto \omega(X, Y)$ induces an isomorphism $v(G) \cong \text{Pic } A(G)$.

The group $\text{Inv } A(G)$ can be computed by using the Mayer-Vietoris sequence of algebraic K -theory and multiplicative congruences for units in the p -adic completion of the Burnside ring.

THEOREM 3 [13]. (i) There exists an exact sequence

$$1 \rightarrow (A(G)/o(G)C(G))^* \rightarrow (C(G)/o(G)C(G))^* \xrightarrow{d} \text{Inv } A(G) \rightarrow 1.$$

Given $z \in \omega(X, Y)$ with values prime to $o(G)$, then $d(z) = \omega(X, Y)$. In particular $\text{Inv } A(G)$ and $v(G)$ are torsion groups.

(ii) Let G be finite. Then $\text{Inv } A(G) \cong \prod (\mathbb{Z}/|WH|)^*$; the product is taken over all $(H) \in \varphi(G)$.

In general, if we fix Y , the G -homotopy types of HRs X with $\text{Dim } X = \text{Dim } Y$ can be specified by suitable degree functions of maps $X \rightarrow Y$.

Let $RO(G)$ and $R(G)$ be the real and complex ring of G , respectively. Then $V \mapsto SV$ induces a homomorphism $l_G: RO(G) \rightarrow V(G)$. Let $TO(G)$ be its kernel and set $JO(G) = RO(G)/TO(G)$. The dimension function of an HR X is called *linear* if there exists a representation V such that $\text{Dim } X = \text{Dim } SV$. It is called *stably linear* if there exist representations V and W such that $\text{Dim } X = \text{Dim } SV - \text{Dim } SW$. Let $1 \rightarrow T \rightarrow G \rightarrow F \rightarrow 1$ be an extension of a torus T by a finite nilpotent group F . The conjugation action of F on the Lie algebra $L(T)$ is called *primary* if its isotypical summands ([4, p. 70]) are lifted from

representations of Sylow subgroups of F . The group G is called *niltoral* if the F -module $L(T)$ is primary.

THEOREM 4 ([9] FOR FINITE, [1] FOR GENERAL G). *All HRs of G have stably linear dimension functions if and only if G is niltoral.*

The stably linear dimension functions in this theorem are exactly those functions $z \in C'(G)$ which satisfy the *Borel-Smith conditions* [24, 2, 3, 14]:

(a) If $H \triangleleft K \subset G$, $K/H \cong \mathbf{Z}/p \times \mathbf{Z}/p$, $H_0/H, \dots, H_p/H \subset K/H$ of index p , then $n(H) - n(K) = \sum_{i=0}^p n(H_i) - n(K)$.

(b) If $H \triangleleft K \subset G$, $K/H \cong \mathbf{Z}/p$, p odd, then $n(H) - n(K)$ is even.

(c) If $H \triangleleft K \triangleleft L \subset G$, $K/H \cong \mathbf{Z}/2$, $L/H \cong \mathbf{Z}/4$ (resp. L/H quaternion 2-group), then $n(H) - n(K)$ is even (resp. divisible by 4).

We want to characterize the linear dimension functions among the stably linear dimension functions. A group G is called *primitive* if all its normal abelian subgroups are cyclic. If $z \in C'(G)$ and $H \triangleleft K \subset G$, then define $z_{K/H} \in C'(K/H)$ by $z_{K/H}(L/H) = z(L)$.

PROPOSITION 5. *Let $z \in C'(G)$ be stably linear. Suppose that, for each primitive subquotient K/H , $z_{K/H}$ is linear. Then z is linear.*

For example, a finite nilpotent group is primitive if and only if its p -Sylow groups are: (i) cyclic, if p is odd; (ii) cyclic, dihedral, quaternion, or semidihedral, if $p = 2$. Let D be cyclic. Then $z \in C(D)$ is linear if and only if $\sum_{E \subset D} \mu(|E|)z(E)$, and similar sums for all subquotients, are nonnegative (μ Möbius function). If X is a locally linear RF of G and $X^H \neq \emptyset$, then X^H , considered as a WH-homotopy representation, is linear [21]. A special case of Proposition 5 is

PROPOSITION 6 [9, 19]. *Let G be a finite p -group. Then HRs of G have linear dimension functions.*

THEOREM 7 [10]. *Let T be a torus. Then the natural map $l_T: RO(T) \rightarrow V(T)$ is an isomorphism.*

In general, the map l_G has a nontrivial kernel and cokernel. For example: If G is finite abelian, then the subgroup of $\text{Inv } A(G)$ consisting of the $\omega(SV, SW)$ corresponds to the factors belonging to $(H) \in \varphi(G)$ with cyclic G/H (see Theorem 3, (iii)). Thus, already for $G = \mathbf{Z}/p \times \mathbf{Z}/p$, there exist nonlinear homotopy types with linear dimension function.

Let G be finite and set $B = \{(H) \in \varphi(G) | H/[H, H] \text{ cyclic}\}$. Then

THEOREM 8 [18]. *The image of $\text{Dim}: V(G) \rightarrow C(G)$ has rank equal to cardinality of B . The set B consists of the cyclic conjugacy classes if and only if G is nilpotent. Thus, $\dim JO(G) \otimes \mathbf{Q} = \dim V(G) \otimes \mathbf{Q}$ if and only if G is nilpotent.*

THEOREM 9 [1]. *$l_G \otimes \mathbf{Q}$ is an isomorphism if and only if G is niltoral.*

Already for cyclic groups, there exist nonlinear dimension functions which are stably linear. In order to describe the dimension functions unstably, we assume that HRs X satisfy two further conditions: (iv) the set $\text{Iso}(X)$ of isotropy subgroups of X is closed under intersections; (v) if $H \in \text{Iso}(X)$ and H is strictly contained in L , then $n(L) < n(H)$.

A G -space X is called **\mathbf{Z} -homology finite** if all fixed point sets X^K have finitely generated homology groups $H_*(X^K; \mathbf{Z})$. Let $\mathbf{Z}_{(p)}$ denote the integers localized at the prime ideal (p) . A space Y is called a **$\mathbf{Z}_{(p)}$ -homology sphere** of (homological) dimension $h\text{-dim } Y = n$ if $H_*(Y; \mathbf{Z}_{(p)}) \cong H_*(S^n; \mathbf{Z}_{(p)})$. Let P be a finite p -group and let Y be a **\mathbf{Z} -homology finite, finite-dimensional P -complex**, which is a **$\mathbf{Z}_{(p)}$ -homology sphere**. Then, by P. A. Smith theory, the fixed point sets Y^L , $L \subset P$, are again **$\mathbf{Z}_{(p)}$ -homology spheres**, of dimension $d(L)$ say. Thus we have a dimension function $\text{Dim } Y: (L) \mapsto d(L) + 1$.

If $H \subset G$ and if Y is an H -space, then a **G -invariance structure** on Y is an H -map $(\lambda_g) = \lambda: G \times_H Y \rightarrow Y, (g, y) \mapsto \lambda_g(y) = \lambda(g, y)$ such that the two maps $G \times_H (G \times_H Y) \rightarrow Y$ given by $(g_1, g_0, x) \mapsto \lambda(g_1, \lambda(g_0, x))$ and $(g_1, g_0, x) \mapsto \lambda(g_1 g_0, x)$ are H -homotopic, and $Y \rightarrow Y, x \mapsto \lambda(1, x)$ is the identity. As a consequence of this definition we have: The map $\lambda_g: Y \rightarrow Y$ is a homotopy equivalence with the following equivariance property: $k\lambda_g(x) = \lambda_g(g^{-1}kgx)$, for $g \in G, x \in Y, k \in H \cap gHg^{-1}$. One should think of λ as an extension of the H -action to a G -action up to homotopy, with λ_g as the left translation by g . Restriction to subgroups induces invariance structures on fixed point sets.

For $D \in C(G)$ we define $\text{Iso}(D) = \{H | D(H) > 0; H \subset K, H \neq K \Rightarrow D(K) < D(H)\}$. With this definition we have: If X is an HR satisfying (iv) and (v), and if $D := \text{Dim } X$, then $\text{Iso}(X) = \text{Iso}(D)$. Let $G(p)$ denote a p -Sylow subgroup of G .

PROPOSITION 10. *Let G be finite. Suppose $D = \text{Dim } X$ is the dimension function of a homotopy representation. Then the function D has the following properties:*

- (i) *For $H \subset K \subset G$ we have $D(H) \geq D(K) \geq 0$.*
- (ii) *$\text{Iso}(D)$ is closed under intersections.*
- (iii) *For each $H \subset G$ and each prime p (dividing the order of WH) there exists a finite-dimensional, \mathbf{Z} -homology finite $WH(p)$ -complex $Y(H, p)$ with WH -invariance structure $(\lambda_w | w \in WH)$; moreover $Y(H, p)$ is a **$\mathbf{Z}_{(p)}$ -homology sphere** and satisfies $D_{WH(p)} = \text{Dim } Y(H, p)$.*
- (iv) *The given invariance structure (λ_w) on $Y(H, p)$ satisfies: For each p -group $P = U/H \subset WH(p)$ and $w \in WH$ with $wPw^{-1} = P$, we have $\text{degree } \lambda_w^P = (-1)^{D(U) - D(w, U)}$. Here (w, U) is the preimage in NH of the group generated by w and U .*

It turns out that these necessary conditions on dimension functions are also essentially sufficient. The construction of homotopy representations uses cell-attaching and the resolution theory of [25, 9, 11, 12]. Therefore we exclude dimensions less than 3.

THEOREM 11. *Let G be finite. Let $D \in C(G)$ be a function having the properties (i)–(iv), listed in Proposition 10. Let $S = \{H \in \text{Iso } D \mid D(H) \leq 3\}$. Let C be a 2-dimensional G -complex such that $\text{Iso}(C) = S$ and C^H is a $D(H)$ -dimensional complex which is homotopy equivalent to $S^{D(H)}$ for $H \in S$. Then there exists a homotopy representation X of G such that $D = \text{Dim } X$, C is a subcomplex of X , and $C^H = X^H$ for $H \in S$.*

We mention a qualitative result which can be obtained from Theorem 11.

PROPOSITION 12. *Let $D \subset C(G)$ satisfy (i) and (ii) of Proposition 10. Suppose D satisfies the Borel-Smith conditions (a), (b), and (c). Then $2\varphi(|G|)D$ is the dimension function of a homotopy representation.*

(Here φ is the Euler function.)

The first assertion of Theorem 8 is essentially a special case of Proposition 12.

In order to obtain explicit calculations of dimension functions from Theorem 11, it is necessary to study the homotopy theory of representations. This is then used to analyze invariance structures on representation spheres. Orthogonal representations V and W of G over \mathbf{R} are called: *homotopy equivalent*, $V \sim W$, if SV and SW are G -homotopy equivalent; *stably homotopy equivalent*, $V \sim_s W$, if, for some U , $V \oplus U \sim W \oplus U$. Unitary representations V and W of G over \mathbf{C} are called: *oriented homotopy equivalent*, $V \sim_o W$, if there exists a G -map $f: SV \rightarrow SW$ such that, for all $H \subset G$, $\text{degree } f^H = 1$; *stably oriented homotopy equivalent*, $V \sim_{so} W$, if, for some unitary U , $V \oplus U \sim_o W \oplus U$.

Let Γ denote the Galois group of the cyclotomic numbers over \mathbf{Q} . Then Γ acts on representations of finite groups; if V is irreducible, then so is V^γ , for $\gamma \in \Gamma$. The following is proved using [26].

PROPOSITION 13. *$V \oplus V^{\alpha\beta} \sim V^\alpha \oplus V^\beta$, for all representations V of the finite group G and all $\alpha, \beta \in \Gamma$ (similarly in the unitary case).*

We define additive subgroups of the representation rings $RO(G)$ and $R(G)$ for finite G as follows:

$$RO_0(G) = \{V - W \mid \text{Dim } V = \text{Dim } W\}$$

$$RO_h(G) = \{V - W \mid V \sim_s W\}$$

$$RO_1(G) = \text{subgroup generated by } V \oplus V^{\alpha\beta} - V^\alpha \oplus V^\beta;$$

and similarly $R_0(G)$, $R_h(G)$, and $R_1(G)$ in $R(G)$ where $R_h(G)$ is defined in terms of oriented equivalence. We have $RO_0(G) \supset RO_h(G) \supset RO_1(G)$, the latter inclusion by Proposition 13.

THEOREM 14. (i) $V \sim_{so} W \Rightarrow V \sim_o W$ [8].

(ii) Suppose G is finite of odd order. Then $V \sim_s W \Rightarrow V \sim W$ [28].

(iii) Suppose $1 \rightarrow N \rightarrow G \rightarrow G/N \rightarrow 1$, G/N supersolvable, N solvable with all Sylow groups abelian. Then $V \sim_s W \Rightarrow V \sim W$ (Tornehave).

(iv) Let G be finite. Suppose $V \sim_s W$. If $V - W \in RO_h(G)/RO_1(G)$ is divisible by 2 in this group, then $V \sim W$ (Tornehave).

By Artin induction one can show that $|G|RO_h(G)/RO_1(G) = 0$. Thus, (iv) \Rightarrow (ii). The general problem of the stable homotopy theory of representations is the computation of $RO_h(G)$ and $R_h(G)$ in algebraic terms. Partial results are known; for nilpotent groups the results are particularly neat.

THEOREM 15 [8]. *Let G be a finite p -group. Then*

(i) $RO_1(G) = RO_h(G)$, $R_1(G) = R_h(G)$.

(ii) *Suppose V and W are irreducible. Then $V \sim W \Rightarrow V \cong W$.*

The equality (i) also holds for nilpotent groups of odd order.

For compact Lie groups of positive dimension the situation is entirely different.

THEOREM 16 [27]. *Let G be a compact Lie group. Suppose $V \sim_s W$ and suppose that V and W contain no irreducible summands U with finite $G/\text{Ker } U$. Then $V \cong W$.*

For certain classes of finite groups, homotopy equivalences can only occur between Galois conjugate representations. If U is an irreducible representation, then let $V(U)$ denote the sum of all subrepresentations of V which are Galois conjugate to U (over \mathbf{R} resp. \mathbf{C}).

THEOREM 17. *Suppose G is as in Theorem 14, (iii). Then $V \sim W \Rightarrow V(U) \sim W(U)$ for all irreducible U .*

In general, there is the following *problem*: Let U be an irreducible $\mathbf{R}G$ -module; determine the *Galois homotopy isotropy group*, i.e., the group $\Gamma(U) = \{\gamma \in \Gamma | U^\gamma \sim U\}$, in algebraic terms! The computations of $\Gamma(U)$, $jO(G) = RO_h(G)/RO_1(G)$, and $j(G) = R_h(G)/R_1(G)$ can be based on the fact that the natural maps $jO(G) \rightarrow \text{Pic } A(G)$, $j(G) \rightarrow \text{Inv } A(G)$ are injective [17]. If $V \sim_o W$, there is a unique G -homotopy equivalence $f: S(V) \rightarrow S(W)$, up to G -homotopy. Its degree in complex G -equivariant K -theory is a unit $a(f) \in R(G)^*$.

PROPOSITION 18. *Let G be finite. The units of $R(G)^*$ which arise as $a(f)$ for homotopy equivalences $f: SV \rightarrow SW$ form a subgroup of finite index.*

Another, largely untouched, problem is: Which HRs are realisable as RFs? For free actions, this is the well-known spherical space form problem. It is shown in [21] that different dimension functions are realisable in PL and DIFF, even for cyclic groups of odd order. Nonlinear dimension functions also exist for S^1 -actions [22, 23].

For linear dimension functions, there still exist nonlinear homotopy types. They can be realized as RFs with exotic linking number for the fixed point sets [15, 16]. The simplest examples occur for $G = \mathbf{Z}/p \times \mathbf{Z}/p$; there exist RFs X with 3 isotropy groups 1, H , and K and such that X^H and X^K have any preassigned linking number d in X which satisfies $d^2 \equiv d \pmod{p^2}$. If $d \not\equiv 1 \pmod{p^2}$, then the homotopy type is nonlinear.

The condition $d^2 \equiv d \pmod{p^2}$ is a finiteness obstruction. There exists a homomorphism

$$s: V(G) \rightarrow \prod_{(H) \in \varphi(G)} \tilde{K}_0(\mathbb{Z}\pi_0 WH)$$

whose kernel is represented by differences of finite homotopy representations. The map s can be computed in terms of Swan homomorphisms $(\mathbb{Z}/|W|)^* \rightarrow \tilde{K}_0(\mathbb{Z}W)$; see [25, 20, 13, 18].

An example for the internal geometry of RFs are *linking systems*. Let V_1, \dots, V_n be representations of G . For $\emptyset \neq \sigma \subset \{1, \dots, n\}$ let $S(\sigma) = S(\bigoplus_{i \in \sigma} V_i)$. An RF X together with a family of G -invariant sub-RFs $X(\sigma)$ and a G -map $f: X \rightarrow S = S(\{1, \dots, n\})$ with $fX(\sigma) \subset S(\sigma)$, such that $f: X(\sigma) \rightarrow S(\sigma)$ is of degree d for all σ , is called a d -fold of the system $(S(\sigma))$. If the X and $X(\sigma)$ are smooth with transverse intersection and if f is a stratified and transverse linear map with respect to the stratifications $(X(\sigma))$ and $(S(\sigma))$, then the linking number of $X(\sigma), X(\mu)$ in $X(\sigma \cup \mu)$ for $\sigma \cap \mu = \emptyset$ is always d . Such configurations are therefore called linking systems.

THEOREM 19. *A d -fold linking system exists under the following hypotheses:*

- (i) G is finite of odd order;
- (ii) $V = V_1 \oplus \dots \oplus V_n$ is a faithful representation, $\dim_{\mathbb{R}} V_i \geq \max(n, 5)$;
- (iii) the number d is contained in the kernel of the Swan homomorphism $(\mathbb{Z}/|G|)^* \rightarrow \tilde{K}_0(\mathbb{Z}G)$.

The proof is by stratified surgery, analogous to [5]. It is conjectured that d -folds of representations $S(V)$ exist quite generally. Linking systems realize many HRs with nonlinear homotopy type.

One of the difficult and unsolved problems in geometric representation theory is the computation of the Grothendieck group of topological spherical RFs with join as composition law. This is the topological analogue of the representation ring.

REFERENCES

1. S. Bauer, *Dimensionsfunktionen von Homotopiedarstellungen kompakter Liescher Gruppen*, Dissertation, Univ. Göttingen, 1986.
2. A. Borel, *Seminar on transformation groups*, Princeton University Press, Princeton, N.J., 1960.
3. G. E. Bredon, *Introduction to compact transformation groups*, Academic Press, New York, 1972.
4. Th. Bröcker and T. tom Dieck, *Representations of compact Lie groups*, Springer, Berlin-Heidelberg-New York, 1985.
5. W. Browder and F. Quinn, *A surgery theory for G -manifolds and stratified sets* (Conf. Proc., Tokyo, 1973), Univ. of Tokyo Press, Tokyo, 1975, pp. 27–36.
6. T. tom Dieck, *The Burnside ring of a compact Lie group*. I, Math. Ann. **215** (1975), 235–250.
7. ———, *A finiteness theorem for the Burnside ring of a compact Lie group*, Compositio Math. **35** (1977), 91–97.
8. ———, *Homotopy-equivalent group representations*, J. Reine Angew. Math. **298** (1978), 192–195.

9. —, *Homotopiedarstellungen endlicher Gruppen: Dimensionsfunktionen*, Invent. Math. **67** (1982), 231–252.
10. —, *Homotopy representations of the torus*, Arch. Math. **38** (1982), 459–469.
11. —, *The singular set of group actions on homotopy spheres*, Arch. Math. **43** (1984), 551–558.
12. —, *The homotopy type of group actions on homotopy spheres*, Arch. Math. **45** (1985), 174–179.
13. —, *The Picard group of the Burnside ring*, J. Reine Angew. Math. **361** (1985), 174–200.
14. —, *Transformation groups*, W. De Gruyter, Berlin-New York, 1986.
15. T. tom Dieck and P. Löffler, *Verschlingung von Fixpunktmenngen in Darstellungsformen*. I (Proc. Conf. Alg. Topology, Göttingen, 1984), Lect. Notes in Math., vol. 1172, Springer-Verlag, Berlin-New York, 1985, pp. 167–187.
16. —, II (Proc. Conf. Alg. Topology, Poznan, 1985), Lect. Notes in Math., Springer-Verlag, Berlin-New York, 1986.
17. T. tom Dieck and T. Petrie, *Geometric modules over the Burnside ring*, Invent. Math. **47** (1978), 273–287.
18. —, *Homotopy representations of finite groups*, Publ. Math. I.H.E.S. **56** (1982), 129–169.
19. R. M. Dotzel and G. Hamrick, *p-group actions on homology spheres*, Invent. Math. **62** (1981), 437–442.
20. W. Lück, *The geometric finiteness obstruction*, J. London Math. Soc., to appear.
21. I. Madsen and M. Raußen, *Smooth and locally linear G-homotopy representations* (Proc. Conf. Alg. Topology, Göttingen, 1984), Lect. Notes in Math., vol. 1172, Springer-Verlag, Berlin-New York, 1985, pp. 130–156.
22. D. Montgomery and C. T. Yang, *Differentiable pseudo-free circle actions on homotopy seven spheres* (Proc. Conf. Trans. Groups, Amherst, 1971), Lect. Notes in Math., vol. 298, Springer-Verlag, Berlin-New York, 1972, pp. 41–101.
23. T. Petrie, *Pseudoequivalences of G-manifolds*, Proc. Sympos. Pure Math., vol. 32, Amer. Math. Soc., Providence, R.I., 1978, pp. 169–210.
24. P. A. Smith, *Transformations of finite period*, Ann. of Math. **39** (1938), 127–164.
25. R. G. Swan, *Periodic resolutions for finite groups*, Ann. of Math. **72** (1960), 267–291.
26. J. Tornehave, *Equivariant maps of spheres with conjugate orthogonal actions* (Proc. Conf. Alg. Topology, London, Ont., 1981), Canadian Math. Soc. Conf. Proc., vol. 2., part 2, 1982, pp. 275–301.
27. P. Traczyk, *On the G-homotopy equivalence of spheres of representations*, Math. Z. **161**, 257–261 (1978).
28. —, *Cancellation law for homotopy equivalent representations of groups of odd order*, Manuscripta Math. **40** (1982), 135–154.

UNIVERSITY OF GÖTTINGEN, GÖTTINGEN, FEDERAL REPUBLIC OF GERMANY

Periods of Abelian Integrals, Theta Functions, and Differential Equations of KdV Type

E. ARBARELLO

The generalized Siegel upper half plane $\mathcal{H}_g \subset \mathbf{C}^{g(g+1)/2}$ is the set of complex $g \times g$ symmetric matrices with positive definite imaginary part. Given a matrix $\tau \in \mathcal{H}_g$ one can construct the complex torus $X_\tau = \mathbf{C}^g / \mathbf{Z}^g + \tau \mathbf{Z}^g$ together with Riemann's theta function

$$\theta(z, \tau) = \sum_{p \in \mathbf{Z}^g} \exp 2\pi i \left\{ \frac{1}{2} {}^t p \tau p + {}^t p z \right\}.$$

Consider now a compact Riemann surface C of genus g . Fix a symplectic basis $a_1, \dots, a_g, b_1, \dots, b_g$ of $H_1(C, \mathbf{Z})$ and a basis $\omega_1, \dots, \omega_g$ of the vector space of homomorphic differentials on C having the property that $\int_{a_i} \omega_j = \delta_{ij}$. As Riemann noticed, the matrix $\tau(C)$ of the "b-periods" is symmetric and has positive imaginary part. The matrices of \mathcal{H}_g that are obtained in this way form a $(3g-3)$ -dimensional analytic subvariety \mathcal{I}_g of \mathcal{H}_g which is called the Jacobian locus. To each point $\tau(C)$ of \mathcal{I}_g corresponds the Jacobian variety $J(C) = X_{\tau(C)}$ of C , together with the theta function $\theta(z, \tau(C))$.

The problem of finding explicit equations for \mathcal{I}_g inside \mathcal{H}_g , goes back to Riemann, Schottky, and Jung [R], [S], [S-J].

In his study of solutions of nonlinear equations of Korteweg de Vries type, Krichever [K] proved the following fact:

Fix a compact Riemann surface \mathbf{C} of genus $g > 0$ and consider its theta function $\theta(z)$, then there exist three vectors $W^{(1)}, W^{(2)}, W^{(3)}$ in \mathbf{C}^g , with $W^{(1)} \neq 0$, such that, for every $z \in \mathbf{C}^g$ the function

$$u(x, y, t; z) = \frac{\partial^2}{\partial x^2} \log \theta(xW^{(1)} + yW^{(2)} + tW^{(3)} + z) \quad (1)$$

satisfies the so-called Kodomcev-Petviashvili equation

$$3u_{yy} = \frac{\partial}{\partial x} (u_t - 3uu_x - 2u_{xxx}). \quad (2)$$

S. P. Novikov [N] then conjectured that a matrix $\tau \in \mathcal{H}_g$, which is not in block form, is a Riemann matrix if and only if the corresponding theta function

satisfies the K.P. equation, in the sense we just explained. A strong indication of the validity of this conjecture was given by Dubrovin in [DU2].

Following work of Mumford [M2] and Mulase [M], Novikov's conjecture has been recently proved by Shiota [S].

A more geometrical approach to the same problem originated from the following "trisecant formula," discovered by Fay [FY]. Fix a compact Riemann surface C of genus $g > 0$. Look at the Abel-Jacobi map of C into its Jacobian and denote by Γ the isomorphic image of C under this map. Fay's trisecant formula says that given any three points α, β, γ on Γ and any point ζ on $\frac{1}{2}(\Gamma - \alpha - \beta - \gamma)$, there exist constants c_1 and c_2 such that

$$\theta(z - \alpha)\theta(z + 2\zeta + \alpha) = c_1\theta(z - \beta)\theta(z + 2\zeta + \beta) + c_2\theta(z - \gamma)\theta(z + 2\zeta + \gamma). \quad (3)$$

As Mumford noticed [M4], when α, β, γ tend to 0 on Γ , one then gets the equation

$$\begin{aligned} D_1^4\theta \cdot \theta - 4D_1^3\theta \cdot \theta + 3(D_1^2\theta)^2 - 3(D_2\theta)^2 \\ + 3D_2^2\theta \cdot \theta + 3D_1\theta \cdot D_3\theta - 3D_1D_3\theta \cdot \theta + d_4\theta \cdot \theta = 0 \end{aligned} \quad (4)$$

where d_4 is a suitable constant and where D_1, D_2, D_3 are the constant vector fields on $J(C)$ obtained as follows. Let $W^{(1)}, W^{(2)}, W^{(3)}$ be, respectively, the tangent, the normal, and the binormal vector to Γ at 0, and set

$$D_i = \sum_{j=1}^g W_j^{(i)} \frac{\partial}{\partial \zeta_j}, \quad i = 1, 2, 3.$$

A computation shows that the K.P. equation (2) is equivalent to (4) via the substitution (1).

One can then express Novikov's conjecture as follows.

(5) THEOREM ([S] AND [AD2]). *Let $X = X_\tau$ be an irreducible principally polarized abelian variety (i.p.p.a.v.) of dimension $g \geq 1$. Then X is the Jacobian of a compact Riemann surface of genus g if and only if there exist a constant d_4 and constant vector fields $D_1 \neq 0, D_2, D_3$ on X such that the equation (4) is satisfied.*

We shall sketch a proof of this theorem following the lines of [AD2]. Let $\Theta \subset X$ be the theta divisor. It is well known that the 2^g functions

$$\hat{\theta}[n](z, \tau) = \sum_{p \in \mathbf{Z}^g} \exp 2\pi i \{ {}^t(p+n)\tau(p+n) + 2 {}^t(p+n)z \}, \quad n \in \tfrac{1}{2}\mathbf{Z}^g/\mathbf{Z}^g,$$

form a basis of $H^0(X, 2\Theta)$ and that, as Riemann discovered,

$$\theta(z + \zeta)\theta(z - \zeta) = \sum_{n \in \frac{1}{2}\mathbf{Z}^g/\mathbf{Z}^g} \hat{\theta}[n](z)\hat{\theta}[n](\zeta). \quad (6)$$

Fay's identity (3) can then be geometrically interpreted in terms of the Kummer map $\vec{\theta}: X \rightarrow \mathbf{P}^N$, $N = 2^g - 1$, given by $\vec{\theta}(\zeta) = [\dots, \hat{\theta}[n](\zeta), \dots]$. In fact when X is a Jacobian, using (6), the trisecant formula becomes

$$\vec{\theta}(\zeta + \alpha) = c_1 \vec{\theta}(\zeta + \beta) + c_2 \vec{\theta}(\zeta + \gamma), \quad \forall \zeta \in \tfrac{1}{2}(\Gamma - \alpha - \beta - \gamma). \quad (7)$$

This says that the points $\vec{\theta}(\zeta + \alpha)$, $\vec{\theta}(\zeta + \beta)$, $\vec{\theta}(\zeta + \gamma)$ lie on a line which is then a trisecant to the Kummer variety $\vec{\theta}(X)$.

For fixed triple of points α, β, γ on *any* principally polarized abelian variety X , Gunning [G] introduces the subscheme $\tilde{V}_{\alpha, \beta, \gamma} \subset X$ defined by

$$\tilde{V}_{\alpha, \beta, \gamma} = \{\zeta \in X: \vec{\theta}(\zeta + \alpha) \wedge \vec{\theta}(\zeta + \beta) \wedge \vec{\theta}(\zeta + \gamma) = 0\},$$

and proves that: *an i.p.p.a.v. X is a Jacobian if and only if there is a positive dimensional component of $2\tilde{V}_{\alpha\beta\gamma}$ passing through the point $\alpha - \beta$ and if there is no complex multiplication mapping $\beta - \alpha$ and $\gamma - \alpha$ into 0.*

Inspired by this criterion and motivated by the relation between the trisecant formula and the K.P. equation, Welters [W2] gives an infinitesimal version of this criterion by substituting the points α, β, γ with a second order germ of curve

$$\varepsilon \rightarrow W^{(1)}\varepsilon + W^{(2)}\varepsilon^2 \quad (8)$$

at the origin in X . Welters's criterion is the following:

An i.p.p.a.v. X is a Jacobian if and only if there exist constant vector fields $D_1 \neq 0, D_2$ on X such that the subscheme

$$\tilde{V} = \{\zeta \in X: \vec{\theta}(\zeta) \wedge D_1 \vec{\theta}(\zeta) \wedge (D_1^2 + D_2) \vec{\theta}(\zeta) = 0\}$$

has a positive dimensional component passing through the origin.

As Welters remarks, the subscheme \tilde{V} coincides at the origin of X , up to second order, with the germ (8). Therefore, the condition of Welters's criterion can be expressed by saying that there exist vectors $W^{(3)}, W^{(4)}, \dots$ such that the formal germ $\zeta(\varepsilon) = \sum_{i=1}^{\infty} W^{(i)}\varepsilon^i$ is contained in \tilde{V} , i.e.,

$$\vec{\theta}(\zeta(\varepsilon)) \wedge D_1 \vec{\theta}(\zeta(\varepsilon)) \wedge \overline{\Delta}_2 \vec{\theta}(\zeta(\varepsilon)) \equiv 0 \quad (\overline{\Delta}_2 = D_1^2 + D_2). \quad (9)$$

Now set

$$D_i = \sum_{j=1}^g W_j^{(i)} \frac{\partial}{\partial \zeta_j}, \quad i = 1, 2, \dots,$$

$$\Delta_s = \Delta_s(D_1, \dots, D_s) = \sum_{i_1 + 2i_2 + \dots + si_s = s} \frac{1}{i_1! \dots i_s!} D_1^{i_1} \dots D_s^{i_s}.$$

A computation contained in [AD1] shows that, up to a renormalization of the D_i 's, (9) is equivalent to the existence of constants d_i , $i = 1, 2, \dots$, with $d_1 = d_2 = d_3 = 0$, such that

$$\left(\sum_{i \geq 0} d_{i+1} \varepsilon^i \right) \vec{\theta}(\zeta(\varepsilon)) + D_1 \vec{\theta}(\zeta(\varepsilon)) - \varepsilon \overline{\Delta}_2 \vec{\theta}(\zeta(\varepsilon)) \equiv 0. \quad (10)$$

Equating to zero the coefficients of ε^s in this relation and using (6), the condition (10) is equivalent to the equations

$$\left[\Delta_s \Delta_1 - \Delta_s \overline{\Delta}_2 + \sum_{i=1}^s d_{i+1} \Delta_{s-i} \right] \theta(z + \zeta) \theta(z - \zeta)|_{\zeta=0} = 0, \quad s \geq 1. \quad (11)$$

The left-hand side of (11) only depends on the knowledge of $D_1, \dots, D_s, d_4, \dots, d_{s+1}$ and we denote it by $P_s(z)$. Clearly $P_1(z)$ and $P_2(z)$ are identically zero and, as expected, the equation $P_3 = 0$ is the K.P. equation. Welters's criterion is then translated in the following (cf. [AD1]).

An i.p.p.a.v. X is a Jacobian if and only if there exist constant vector fields $D_1 \neq 0, D_2, D_3, \dots$ and complex numbers d_4, d_s, \dots such that $P_s(z) = 0, s \geq 1$. (12)

(It is interesting to notice that equations (11) are all part of the so-called K.P. hierarchy (see [D] and [AD1]).

The next step in the proof of (5) consists in restricting equation (11) to the subscheme $D_1\Theta \subset X$ defined by the simultaneous vanishing of θ and $D_1\theta$. The first observation is that $P_s(z)$ is a section of $\mathcal{O}_X(2\Theta)$. This section can be written as

$$P_s(z) = P'_s(z) + 2D_1D_s\theta \cdot \theta - 2D_1\theta \cdot D_s\theta + d_{s+1}\theta \cdot \theta$$

where $P'_s(z)$ only depends on the knowledge of $D_1, \dots, D_{s-1}, d_4, \dots, d_s$. Looking at the cohomology sequences given by restricting $\mathcal{O}(2\Theta)$ to Θ and $D_1\Theta$, one can prove the following statement. Given $D_1, \dots, D_{s-1}, d_4, \dots, d_s$, there exist D_s and d_{s+1} such that $P_s(z) = 0$ if and only if $P'_s|_{D_1\Theta} = 0$ or, equivalently, if and only if one can find a meromorphic solution in \mathbf{C}^g to the equation

$$D_1h = P'_s/\theta^2. \quad (13)$$

The existence of D_s and d_{s+1} is thus translated into analytical terms. Via a cohomological argument one then shows that to find a global solution of (13) it is in fact sufficient to find local solutions near every point of

$$U = \mathbf{C}^g \setminus \{z \in \mathbf{C}^g : \theta(z) = D_1\theta(z) = 0\}.$$

The solution of the local problem is easy. First of all one sees that the local integrability condition is given by $\Gamma(P'_s/\theta^2) = 0$ where

$$\Gamma = \theta^4 \left(\frac{1}{3}D_1^4 + D_2^2 - D_1D_3 + 4D_1(D_1^2 \log \theta)D_1 \right),$$

then a direct computation shows that the integrability condition holds under the assumption that $P_3 = \dots = P_{s-1} = 0$. The conclusion is that, under this assumption, there exists D_s and d_{s+1} such that $P_s = 0$. The theorem follows, by induction, from (12).

REFERENCES

- [AD1] E. Arbarello and C. De Concini, *On a set of equations characterizing Riemann matrices*, Ann. of Math. **120** (1984), 119–140.
- [AD2] —, *Another proof of a conjecture of S. P. Novikov on periods of abelian integrals on Riemann surfaces*, preprint.
- [BD] A. Beauville and O. Debarre, *Une relation entre deux approches du problème de Schottky*, preprint.
- [D] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, *Transformation groups for soliton equations*, Proceedings RIMS Symposium Non-Linear Integrable Systems, Classical Theory and Quantum Theory, World Scientific, Singapore, 1983.

- [DU1] B. A. Dubrovin, *Theta functions and non-linear equations*, Russian Math. Surveys **36** (1981), 11–92.
- [DU2] —, *The Kadomcev-Petviashvili equation and the relations between the periods of holomorphic differentials on Riemann Surfaces*, Math. USSR-Izv. **19** (1982), 285–296.
- [F] H. M. Farkas, *On Fay's trisecant formula*, J. Analyse Math. **44** (1984), 205–217.
- [FY] J. D. Fay, *Theta Functions on Riemann Surfaces*, Lectures Notes in Math., Vol. 352, Springer-Verlag, Berlin, Heidelberg, New York, 1973.
- [vG] G. van der Geer, *The Schottky Problem* (Arbeitstagung, Bonn, 1984), Lecture Notes in Math., Vol. 1111, Springer-Verlag, Berlin and New York, 1984.
- [G] R. C. Gunning, *Some curves in Abelian varieties*, Invent. Math. **66** (1982), 377–389.
- [K] I. M. Krichever, *Methods of algebraic geometry in the theory of nonlinear equations*, Russian Math. Surveys **32** (1977), no. 6, 185–213.
- [I] J. Igusa, *Theta Functions*, Grundlehren Math. Wiss. **194** (1972).
- [M] M. Mulase, *Cohomological structure in soliton equations and Jacobian varieties*, J. Differential Geom. **19** (1984), 403–430.
- [M1] D. Mumford, *Curves and their Jacobians*, Univ. of Michigan Press, Ann Arbor, Michigan, 1975.
- [M2] —, *An algebro-geometric construction of commuting operators and of solutions to the Toda lattice equation, Korteweg de Vries equation and related non-linear equations* (Proc. Internat. Sympos. Algebraic Geom., Kyoto, 1977), Kinokuniya Book Store, Tokyo, 1978, pp. 115–153.
- [M3] —, *Tata Lectures on theta*. II, Birkhäuser, Boston, 1983.
- [M4] D. Mumford and J. Fogarty, *Geometric Invariant theory*, Ergeb. Math. Grenzgeb. **34** (1982).
- [N] S. P. Novikov, *The periodic problem for the Korteweg-de Vries equation*, Functional Anal. Appl. **8** (1974), 236–246.
- [R] B. Riemann, *Théorie des fonctions abeliennes*, Oeuvres de Riemann, p. 89 (Journal de Crelle t. 54; 1857).
- [S] F. Schottky, *Zür Theorie der Abelschen Funktionen von vier Variabeln*, J. Reine Angew. Math. **102** (1888), 304–352.
- [S-J] F. Schottky and H. Jung, *Neue Sätze über Symmetralfunktionen und die Abel'schen Funktionen der Riemann'schen Theorie*, S-B Preuss. Akad. Wiss., Berlin, Phys. Math. Kl. **1** (1909), 282–297.
- [S] T. Shiota, *Characterization of Jacobian varieties in terms of soliton equations*, Invent. Math. **83** (1986), 333–382.
- [W1] G. E. Welters, *A criterion for Jacobian varieties*, Ann. of Math. **120** (1984), 497–504.
- [W2] —, *A characterization of non-hyperelliptic Jacobi varieties*, Invent. Math. **71** (1983), 437–440.
- [W3] —, *On flexes of the Kummer variety*, Nederl. Akad. Wetensch. Proc. Ser. A **45** (1983), 501–520.

L'approche géométrique du problème de Schottky

ARNAUD BEAUVILLE

1. Introduction. Le problème de Schottky est la recherche de caractérisations des jacobiniennes parmi toutes les variétés abéliennes. Plus précisément, soit $\mathcal{A}_g (= H_g / \mathrm{Sp}(2g, \mathbb{Z}))$ l'espace des modules des variétés abéliennes principalement polarisées (complexes) de dimension g . Les jacobiniennes forment une sous-variété J_g de \mathcal{A}_g , et il s'agit de trouver des équations de J_g (ou plutôt de son adhérence \overline{J}_g) dans \mathcal{A}_g .

Ce problème a d'abord été étudié d'un point de vue analytique: on cherche à écrire des équations explicites de \overline{J}_g dans \mathcal{A}_g , définies par des formes modulaires sur H_g —par exemple des polynômes en les thêta-constantes $\theta[\frac{m}{n}](0, \tau)$. Le premier résultat dans cette direction est dû à Schottky [S], qui mit en évidence une équation simple satisfaite par les thêta-constantes d'une jacobienne en genre 4. Le fait que cette équation caractérise les jacobiniennes n'a été prouvé que récemment par Igusa ([I], cf. aussi [F]).

Vingt ans plus tard, en utilisant les variétés de Prym, Schottky et Jung donnèrent un procédé systématique pour écrire des polynômes en les thêta-constantes s'annulant sur J_g , à partir d'identités satisfaites par les thêta-constantes générales [S-J]. Van Geemen a démontré que \overline{J}_g est une composante irréductible de la sous-variété de \mathcal{A}_g définie par ces équations [vG]. Malheureusement on ne sait pas écrire explicitement l'ensemble de ces équations.

Dans cet exposé je vais considérer l'approche géométrique de ce problème. Ici il s'agit avant tout d'obtenir des *caractérisations géométriques* des jacobiniennes, conduisant si possible à des équations plus ou moins explicites. Avant d'entrer dans le sujet, je voudrais mentionner que ce type de caractérisations a des applications à d'autres domaines de la géométrie algébrique, par exemple aux questions de rationalité. Soit en effet X une variété projective et lisse de dimension 3, satisfaisant à $H^0(X, \Omega_X^3) = 0$. On associe à X une variété abélienne principalement polarisée, la jacobienne intermédiaire JX ; si celle-ci n'appartient pas à \overline{J}_g , la variété X ne peut être rationnelle (*critère de Clemens-Griffiths* [C-G]). On déduit de ce critère de nombreux exemples de variétés unirationnelles non rationnelles (cf. [B2]), et aussi un exemple de variété non rationnelle dont le produit avec un espace projectif est rationnel [B-CT-S-SD].

traduit géométriquement comme suit: pour tout $\zeta \in A$ tel que $2\zeta = x + y$, les points $\psi(\zeta)$, $\psi(\zeta - a)$ et $\psi(\zeta - x)$ de \mathbf{P}^N sont alignés; autrement dit,

(iv) *La variété de Kummer $K(A, \Theta)$ admet une trisécante.*

Il est facile de voir que les conditions (ii) à (iv) sont équivalentes, et entraînent (i). Ces conditions sont en fait étroitement reliées à la condition d'Andreotti-Mayer:

THÉORÈME [B-D1]. (1) *Les conditions équivalentes (ii) à (iv) entraînent $(A, \Theta) \in \mathcal{N}_{g-4}$ (c'est-à-dire $\dim \text{Sing } \Theta \geq g - 4$);*

(2) *La condition (i) entraîne $(A, \Theta) \in \mathcal{N}_{g-4}$ ou $(A, \Theta) \in \mathcal{A}_{1,g-1}^{(2)}$ (cf. §2).*

Compte tenu du théorème d'Andreotti-Mayer, on obtient aussitôt:

COROLLAIRE. $\overline{\mathcal{J}}_g$ est une composante de l'ensemble des $(A, \Theta) \in \mathcal{A}_g$ vérifiant l'une des conditions (i) à (iv).

Ainsi l'existence d'une trisécante pour $K(A, \Theta)$ entraîne $(A, \Theta) \in \mathcal{N}_{g-4}$. Il est intéressant de tester les conditions (i) et (ii) sur les composantes de \mathcal{N}_{g-4} décrites au §2. Pour toutes ces composantes sauf $\mathcal{E}_{g,0}$, les variétés abéliennes correspondantes possèdent la propriété (i); par contre elles ne vérifient pas (ii), au moins génériquement—à l'exception bien sûr des jacobiniennes. Ceci conduit à conjecturer que la condition (ii) caractérise les jacobiniennes (*conjecture de la trisécante*). Des résultats partiels ont été obtenus dans cette direction, d'abord par Gunning [G], puis par Welters [W1, W2, W3]: ces auteurs caractérisent les jacobiniennes par l'existence d'une famille assez grande de trisécantes. Citons par exemple le résultat de [W2]: si $K(A, \Theta)$ admet une famille de dimension un de trisécantes et si $\dim \text{Sing } \Theta = g - 4$, alors (A, Θ) est une jacobienne. Mentionnons aussi une question voisine: l'existence d'une sous-variété de A de classe $\Theta^2/2$ dans $H^4(A, \mathbf{Z})$ entraîne-t-elle que (A, Θ) est une jacobienne? La réponse n'est connue que pour $g = 4$ [R].

Convenablement énoncées, les conditions (i) à (iv) ci-dessus gardent un sens lorsque les points a , puis x , puis y deviennent infiniment proches de 0: si par exemple a est un vecteur tangent à l'origine, on définit $\Theta \cap \Theta_a$ par les équations $\theta = D_a \theta = 0$. Le point ultime de cette spécialisation est la condition:

(ii') *Il existe un vecteur non nul $a \in T_0(A)$, et deux sections non nulles de $H^0(\Theta \cap \Theta_a, \mathcal{O}(\Theta))$ dont le produit (dans $H^0(\Theta \cap \Theta_a, \mathcal{O}(2\Theta))$) est nul.*

Lorsque (A, Θ) est irréductible, G. Welters a remarqué que cette condition équivaut à dire que la fonction θ satisfait une équation aux dérivées partielles non linéaire, l'équation $K - P$. On voit ainsi facilement; en spécialisant l'inclusion $\Theta \cap \Theta_{p-q} \subset \Theta_{p-s} \cup \Theta_{s-q}$, que la fonction θ d'une jacobienne satisfait à l'équation $K - P$ (cette propriété a été découverte par Fay). Inversement, toute variété abélienne principalement polarisée irréductible satisfaisant à (ii') est une jacobienne: c'est la conjecture de Novikov, prouvée par Shiota [Sh]. En s'appuyant sur les résultats de Gunning et Welters évoqués ci-dessus, Arbarello et De Concini ont obtenu une démonstration géométrique de cette conjecture (cf. l'exposé

d'Arbarello dans ce volume). La version infinitésimale de la conjecture de la trisécante est donc démontrée.

Avant de quitter le problème de Schottky usuel je voudrais mentionner l'article [vG-vG], qui est lié à ce qui précède, et d'autre part l'approche tout-à-fait différente qui s'appuie sur le fait que le diviseur Θ d'une jacobienne est doublement de translation; je renvoie à [L] pour un exposé complet de ce point de vue classique.

4. Le problème de Schottky pour les variétés de Prym. Soient \tilde{C} , C deux surfaces de Riemann compactes, $\pi: \tilde{C} \rightarrow C$ un revêtement étale double, et σ l'involution correspondante de C . La variété de Prym (P, Θ) associée à π est une variété abélienne principalement polarisée définie comme suit [M]: P est l'image de l'endomorphisme $1 - \sigma^*$ de $J\tilde{C}$, et on prouve qu'un translaté convenable du diviseur thêta de $J\tilde{C}$ découpe sur P le diviseur 2Θ .

Les variétés de Prym forment dans \mathcal{A}_g une sous-variété fermée irréductible \mathcal{P}_g de dimension $3g$ pour $g \geq 5$, contenant J_g (il faut pour cela accepter les variétés de Prym associées à des courbes stables définies dans [B1]). Après les jacobiniennes, les variétés de Prym sont certainement les variétés abéliennes principalement polarisées les mieux connues; il est naturel d'essayer de les caractériser, soit par des équations dans \mathcal{A}_g , soit par des propriétés géométriques. Je ne connais pas d'équivalent à l'approche analytique évoquée au §1. Par contre l'approche géométrique décrite aux §2 et 3 s'étend aux variétés de Prym; le parallélisme est d'ailleurs extrêmement frappant, et reste pour moi assez mystérieux.

Soit (P, Θ) une variété de Prym. La définition ci-dessus permet de décrire assez précisément les singularités du diviseur Θ [M, B1]. On en déduit qu'on a $\dim \text{Sing } \Theta \geq g - 6$, avec égalité lorsque (P, Θ) est assez générale et $g \geq 6$ [W4, D1]. Debarre vient de démontrer l'équivalent du théorème d'Andreotti-Mayer [D2]: pour $g \geq 7$, \mathcal{P}_g est une composante irréductible de \mathcal{N}_{g-6} . D'autres composantes de \mathcal{N}_{g-6} sont construites dans [D1].

D'autre part, pour p, q dans \tilde{C} , notons $[p, q]$ l'élément $p + q - \sigma p - \sigma q$ de P . On déduit facilement de la description géométrique du diviseur Θ l'inclusion

$$\Theta \cap \Theta_{[p,q]} \cap \Theta_{[p,r]} \subset \Theta_{[p,s]} \cup \Theta_{[q,r]}$$

pour p, q, r, s distincts dans \tilde{C} , avec $\sigma p \neq q, r$.

Ceci nous amène à considérer, pour une variété abélienne principalement polarisée (A, Θ) , les conditions suivantes, calquées sur celles du §3, et qui sont satisfaites pour les variétés de Prym:

(i) Il existe des éléments distincts non nuls a, b de A tels que l'intersection $\Theta \cap \Theta_a \cap \Theta_b$ ne soit pas entière.

(ii) Il existe des éléments distincts non nuls a, b, x, y de A tels qu'on ait (schématiquement) $\Theta \cap \Theta_a \cap \Theta_b \subset \Theta_x \cup \Theta_y$.

(iii) Il existe a, b, x, y comme ci-dessus, et des scalaires λ, μ, ν, ρ non tous nuls, tels qu'on ait l'identité² (en posant $s = x + y$):

$$\lambda\theta(z)\theta(z-s) + \mu\theta(z-a)\theta(z+a-s) + \nu\theta(z-b)\theta(z+b-s) + \rho\theta(z-x)\theta(z-y) = 0.$$

(iv) La variété de Kummer $K(A, \Theta)$ admet un plan quadrisécant (i.e. coupant $K(A, \Theta)$ en quatre points distincts).

PROPOSITION [B-D2]. Pour $g \geq 7$, \mathcal{P}_g est une composante irréductible de l'ensemble des $(A, \Theta) \in \mathcal{A}_g$ vérifiant l'une des conditions (i) à (iv).

Le résultat est en fait un peu plus précis: nous montrons que lorsque A est simple et a d'ordre infini, chacune des conditions (i) à (iv) entraîne $(A, \Theta) \in \mathcal{N}_{g-6}$. On prouve ensuite que sur une variété de Prym générique (P, Θ) l'intersection $\Theta \cap \Theta_a \cap \Theta_b$ est toujours intègre lorsque a et b sont de torsion; la proposition en résulte.

L'existence d'un plan quadrisécant ne suffit pas ici à caractériser les variétés de Prym: les éléments (A, Θ) de $\mathcal{A}_{1,g-1}^{(2)}$ possèdent aussi cette propriété.

Contrairement au cas des jacobiniennes, on ne peut spécialiser à la fois $[p, q]$, $[p, r]$, $[p, s]$ et $[q, r]$ en 0 lorsque la courbe C est lisse. Cela devient possible lorsque C est singulière. Avec G. Welters, nous avons observé que l'on obtient ainsi, pour la fonction θ d'une variété de Prym associée à des courbes stables singulières, une équation aux dérivées partielles non linéaire, dite équation *BKP* (cf. par exemple [J-M]). Novikov a conjecturé que cette équation caractérise les variétés de Prym de courbes singulières; une démonstration de cette conjecture a été annoncée par Shiota. Nous espérons que la méthode des plans quadrisécants permettra de comprendre géométriquement cette conjecture.

BIBLIOGRAPHIE

[A-M] A. Andreotti et A. Mayer, *On period relations for abelian integrals on algebraic curves*, Ann. Scuola Norm. Sup. Pisa (3) **21** (1967), 189–238.

[B1] A. Beauville, *Prym varieties and the Schottky problem*, Invent. Math. **41** (1977), 149–196.

[B2] —, *Variétés rationnelles et unirationnelles*, Algebraic Geometry—Open Problems, Lecture Notes in Math., vol. 997, Springer-Verlag, Berlin and New York, 1983, pp. 16–33.

[B-CT-S-SD] A. Beauville, J.-L. Colliot-Thélène, J.-J. Sansuc, et P. Swinnerton-Dyer, *Variétés stablement rationnelles non rationnelles*, Ann. of Math. **121** (1985), 283–318.

[B-D1] A. Beauville et O. Debarre, *Une relation entre deux approches du problème de Schottky*, Invent. Math. **86** (1986), 195–207.

[B-D2] —, *Sur le problème de Schottky pour les variétés de Prym* (à paraître).

[C-G] H. Clemens et P. Griffiths, *The intermediate Jacobian of the cubic threefold*, Ann. of Math. **95** (1972), 281–356.

[D1] O. Debarre, *Sur les variétés abéliennes dont le diviseur Θ est singulier en codimension 3* (à paraître).

[D2] —, *Variétés de Prym et ensembles d'Andreotti-Mayer* (à paraître).

[Do] R. Donagi, *The tetragonal construction*, Bull. Amer. Math. Soc. (N.S.) **4** (1981), 181–185.

²Il faut modifier cette équation lorsque $s = a, b$ ou $a + b$.

- [F] E. Freitag, *Die Irreduzibilität der Schottkyrelation (Bemerkung zu einem Satz von J. Igusa)* Arch. Math. (Basel) **40** (1983), 255–259.
- [G] R. Gunning, *Some curves in abelian varieties*, Invent. Math. **66** (1982), 377–389.
- [vG] B. van Geemen, *Siegel modular forms vanishing on the moduli space of curves*, Invent. Math. **78** (1984), 329–349.
- [vG-vG] B. van Geemen et G. van der Geer, *Kummer varieties and the moduli spaces of abelian varieties*, Amer. J. Math. **108** (1986).
- [I] J. Igusa, *On the irreducibility of Schottky's divisor*, J. Fac. Sci. Univ. Tokyo **28** (1981), 531–545.
- [J-M] M. Jimbo et T. Miwa, *Solitons and infinite dimensional Lie algebras*, Publ. Res. Inst. Math. Sci. **19** (1983), 943–1001.
- [L] J. Little, *Translation manifolds and the converse of Abel's theorem*, Compositio Math. **49** (1983), 147–171.
- [M] D. Mumford, *Prym varieties I*, Contributions to Analysis, Academic Press, New York, 1974, pp. 325–350.
- [R] Z. Ran, *On subvarieties of abelian varieties*, Invent. Math. **62** (1981), 459–479.
- [S] F. Schottky, *Zur Theorie der Abelschen Funktionen von vier Variabeln*, J. Reine Angew. Math. **102** (1888), 304–352.
- [S-J] F. Schottky et H. Jung, *Neue Sätze über Symmetrifunktionen und die Abel'schen Funktionen der Riemann'schen Theorie*, S-B. Preuss. Akad. Wiss. (Berlin), Phys. Math. Kl. **1** (1909), 282–297.
- [Sh] T. Shiota, *Characterization of Jacobian varieties in terms of soliton equations*, Invent. Math. **83** (1986), 333–382.
- [W1] G. Welters, *On flexes of the Kummer variety*, Indag. Math. **45** (1983), 501–520.
- [W2] —, *A characterization of non-hyperelliptic Jacobi varieties*, Invent. Math. **74** (1983), 437–440.
- [W3] —, *A criterion for Jacobi varieties*, Ann. of Math. **120** (1984), 497–504.
- [W4] —, *A theorem of Gieseker-Petri type for Prym varieties*, Ann. Sci. École Norm. Sup. **18** (1985), 671–683.
- [We] A. Weil, *Zum Beweis des Torellischen Satzes*, Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl. **2a** (1957), 33–53.

UNIVERSITÉ DE PARIS-SUD, MATHÉMATIQUES-BÂTIMENT 425, 91405 ORSAY CÉDEX, FRANCE

Curves on Higher-Dimensional Complex Projective Manifolds

HERBERT CLEMENS

0. Introduction. Let X be a smooth, irreducible complex submanifold of \mathbf{CP}^N . One of the simplest invariants associated to X is its *canonical line bundle*

$$K = \det T^{1,0}(X),$$

where $T^{1,0}$ denotes the holomorphic cotangent bundle. This line bundle is called positive (or ample) if the sections of some tensor power of it embed X into projective space. K is called negative if K^{-1} is positive. The theme of this presentation will be: The more positive K is, the harder it is to find curves of low genus and/or low degree on X .

1. The case K negative. The fundamental result here is a celebrated theorem of S. Mori [7]:

(1.1) THEOREM. *If K is negative, then there exists a nontrivial mapping*

$$f: \mathbf{P}^1 \rightarrow X.$$

Suppose, in fact, that f is locally immersive. Then the exact sequence

$$0 \rightarrow T(\mathbf{P}^1) \rightarrow f^*T(X) \rightarrow N_{f,X} \rightarrow 0 \quad (1.2)$$

tells us that the normal bundle $N_{f,X}$ to the mapping f has a decomposition

$$N_{f,X} = \mathcal{O}(a_1) \oplus \cdots \oplus \mathcal{O}(a_{n-1}),$$

where $n = \dim X$ and $\sum a_i = -(\deg f^*K) - 2$. Thus, if $n \geq 3$, the rational curve f must move on X , at least to first order. In fact, in all examples that I know, once the degree of the mapping f is sufficiently high, the deformations of f actually cover X . This last is sufficient to verify the generalized Hodge conjecture in the case $n = 3$, a result essentially due to Conte and Murre [5].

There is a lot of interesting "classical geometry" connected with covering families of rational curves on threefolds with negative canonical bundle (called Fano threefolds). I would like to briefly outline what happens in one interesting case.

We let

$$\pi: V \rightarrow \mathbf{P}^3 \quad (1.3)$$

be a branched double covering whose branch locus is a quartic surface B with six nodes (in general position in \mathbf{P}^3). We will call V a “quartic double solid.” Let \tilde{V} be the standard resolution of V obtained by blowing up its nodes. It turns out that

$$J(\tilde{V}) = H^{2,1}(\tilde{V})^*/H_3(\tilde{V})_{\mathbf{Z}}, \quad (1.4)$$

the “intermediate Jacobian of \tilde{V} ” is a principally polarized abelian variety of dimension four. In fact, the generic abelian variety of dimension four is so obtained.

If we take as our point of departure a desire to study the geometry of abelian varieties of dimension four and their moduli space, the above is very useful, since (1.4) gives a geometric realization of “most” (principally polarized) abelian varieties of dimension four. To see one example of how this helps, we consider an algebraic family of curves on \tilde{V} , namely, the family $\{C_f\}_{f \in F}$ obtained by naturally compactifying the family of smooth conics on V which do not meet the nodes of V . It can be shown that F contains curves of the form

$$C_0 = \pi^{-1}(\text{line}),$$

where the line is (once) tangent to the branch locus B , and that the family F is irreducible. We map

$$\begin{aligned} \phi: F &\rightarrow J(\tilde{V}), \\ \{C\} &\mapsto \int_{C_0}^C, \end{aligned} \quad (1.5)$$

the so-called “Abel-Jacobi mapping.”

A degeneration argument, together with the symmetry of the image of ϕ , shows that

$$\phi(F) \in |2\theta|,$$

where $|2\theta|$ denotes the linear system obtained by squaring the linear system of a symmetric theta divisor of $J(\tilde{V})$.

It is not hard to see that all curves of the form

$$\pi^{-1}(\text{line tangent to } B) \quad (1.6)$$

are mapped to the origin of $J(\tilde{V})$, thereby creating a singularity on $\phi(F)$. One can localize the computation of the derivative of ϕ at a curve (1.6) to the singular point of that curve [3]. This allows one to compute the tangent cone $T(\phi(F))_0$ to $\phi(F)$ at the origin of $J(\tilde{V})$:

$$T(\phi(F))_0 = \eta(B), \quad (1.7)$$

where

$$\eta: \mathbf{P}^3 \rightarrow \mathbf{P}^3 \quad (1.8)$$

is the branched double cover given by the linear system of quadrics passing through the six nodes of B . (The identification of the left-hand and right-hand sides of (1.7) is obtained using the identification

$$\begin{aligned} \text{quadrics through nodes} &\leftrightarrow H^{2,1}(\tilde{V}), \\ Q &\mapsto \text{Residue } Q\Omega/F^{3/2}, \end{aligned}$$

where F defines B and $\Omega = \sum (-1)^i X_i dX_0 \cdots \hat{d}X_i \cdots dX_3$.)

The $\phi(F)$ has a fourth-order singularity at 0 so that

$$\phi(F) \in |2\theta|_{00} \cong \mathbf{P}^4, \tag{1.9}$$

the linear subseries consisting of divisors with at least a fourth-order singularity at 0. So the fibre of the mapping

$$\begin{array}{ccc} \text{moduli space of} & \rightarrow & \text{moduli space of} \\ \text{quartic double solids} & & \text{4-dimensional} \\ \text{with six nodes} & & \text{abelian varieties} \end{array} \tag{1.10}$$

is obtained by picking out of

$$\{T(D)_0 : D \in |2\theta|_{0,0}\}$$

those quartics occurring as images of quartics with six nodes under mappings (1.8). Unpublished work of R. Donagi leads one to a conjecture about the hypersurface in the \mathbf{P}^4 in (1.9) formed by taking all these “geometrically given” quartics. Namely, one conjectures that this hypersurface is simply the cubic threefold which Donagi associates to the four-dimensional principally polarized abelian variety $J(\tilde{V})$.

All this is in close analogy with the classically studied mapping

$$C \times C \rightarrow J(C), \quad (p, q) \mapsto \int_p^q \tag{1.11}$$

for a curve C of genus 3. The image of the mapping (1.11) is the unique element of $|2\theta|_{00}$ in that case. In particular, when our $J(\tilde{V})$ acquires one vanishing theta-null, the elements $\phi(F)$ in $|2\theta|_{00}$ come to twice cover the theta divisor associated with the theta-null, just as happens when C becomes hyperelliptic. Algebraically what is happening is that the equation F for B is becoming quadratic in the quadrics through the singular points of B .

2. The case K trivial. We have seen that projective manifolds X whose canonical divisors are negative tend to have covering families of rational curves, often with very nice properties. These curves tend to parametrize the cohomology of X in geometrically interesting ways.

The situation changes dramatically when the canonical bundle K is trivial. In dimension 3, for example, when

$$H^{3,0}(X) \neq 0$$

the intermediate Jacobian becomes

$$J(X) = (H^{3,0}(X) \oplus H^{2,1}(X))^* / H_3(X)_{\mathbf{Z}}, \quad (2.1)$$

a complex torus but not an abelian variety in any fully satisfactory way. Also the image of any Abel-Jacobi mapping $\phi: F \rightarrow J(X)$, with F connected, can span at most a subtorus annihilated by $H^{3,0}(X)$. The generalized Hodge conjecture is the statement that any such subtorus is parametrized by some Abel-Jacobi mapping ϕ . So curves on X can parametrize at most a countable union of proper subvarieties of $J(X)$. We formulate things as follows:

$$\begin{aligned} G(X) &= \bigoplus \{ \mathbf{Z} \cdot C : C \text{ an irred. alg. curve on } X \}, \\ &\text{equiv. rel. generated by making } C \text{ and } C' \\ &\text{equivalent if they lie in a family } F \text{ of} \\ &\text{curves, where } F \text{ is irreducible and rational,} \\ \mathcal{H}(X) &= \{ \alpha \in G(X) : \alpha \sim 0 \text{ in } H_2(X)_{\mathbf{Z}} \}. \end{aligned}$$

We then have an Abel-Jacobi mapping

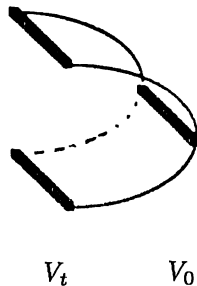
$$\begin{aligned} \eta: \mathcal{H}(X) &\rightarrow J(X), \\ \alpha &\mapsto \int_{\Gamma}, \end{aligned} \quad (2.2)$$

where Γ is a 3-chain with $\partial\Gamma = \alpha$. In all examples with K trivial which have been computed, the image group

$$\eta(\mathcal{H}(X))$$

is dense in $J(X)$. Also, for several examples the image group is a totally disconnected group of infinite rank. (See, for example, [8, 9, or 4].)

One's ability to effectively compute the derivative of the Abel-Jacobi mapping is important in obtaining these results. As an example, imagine a family $\{V_t\}$ of smooth hypersurfaces of degree five in \mathbf{P}^4 . ($K = 0$ for all the V_t .) Suppose that V_t contains two lines which come together at $t = 0$:



Let \tilde{t} parametrize a double cover of the t -line branched at $t = 0$. Then the family of lines under study is parametrized by \tilde{t} . The normal bundle to the line L_0 in

V_0 has a section ν corresponding to the fact that L_0 stays inside V_0 to first order. The derivative of the Abel-Jacobi mapping

$$\text{double cover of } t\text{-space} \rightarrow U_t J(V_t),$$

$$\tilde{t} \rightarrow \int_{\text{base curve in } V_t}^{L_i}$$

is given by the contraction mapping

$$H^0(N_{L_0, V_0}) \otimes H^1(\Omega_{V_0}^2) \rightarrow H^1(\Omega_{V_0}^1) \simeq \mathbb{C}. \quad (2.3)$$

The computation of this derivative can be reduced to an easy residue computation at the four singular points of a plane section of V_0 containing L_0 , and is easily shown to be nonzero. Thus the images of the two lines of V_t in $J(V_t)$ are moving apart near $t = 0$. So the difference of these lines is generically of infinite order in $\mathcal{H}(V_t)$, since this difference is not zero in $\mathcal{H}(V_t)$ for t near 0 but the difference goes to 0 at $t = 0$.

There are some cases where K is trivial, so that the mapping η in (2.2) cannot be surjective, yet the Hodge conjecture predicts that a subtorus of $J(X)$ must lie in the image of η . For example, if η has an involution leaving the sections of K invariant, then the skew-symmetric part of $J(X)$ is conjectured to lie in the image of η . F. Bardelli has verified this for the family of intersections of four quadrics in \mathbb{P}^7 admitting a certain type of involution [1].

Notice that, in the above example, the line L_0 does not deform to first order with the first-order deformation of V_0 in the direction V_t . In fact, a surprisingly easy fact is that for a threefold X with trivial canonical bundle, the “infinitesimal Abel-Jacobi mapping” (2.3) is nontrivial at a (smooth) curve C if and only if there is a first-order deformation of X in which C does not deform.

3. The case K positive. This is the case in which our manifold X is of “general type.” Roughly speaking, when K is positive, it is much harder to find interesting curves of low degree or genus on X . As was mentioned in the K trivial case, once

$$H^{3,0}(X) = H^0(X; K) \neq 0$$

for threefolds, we can't expect curves to parametrize much of $J(X)$. An interesting bit of evidence in favor of the Hodge conjecture comes from studying examples in which K is ample yet $H^{3,0}(X) = 0$. The generalized Hodge conjecture predicts in this case that the Abel-Jacobi mapping (2.2) is surjective. M. Reid suggested studying one such example, the intersection of three sextics in the weighted projective space $\mathbb{P}(2^4, 3^3)$. D. Ortland has recently shown that indeed the Abel-Jacobi mapping (2.2) is surjective in that case, using curves of genus 2 to parametrize $J(X)$.

We should justify our statement about the difficulty of finding interesting curves in the presence of an ample or positive canonical bundle. Actually we can only say something in the case in which X is a generically chosen hypersurface

of some degree m in \mathbf{P}^n . In this case

$$K \approx \mathcal{O}_X(m - (n + 1))$$

so that we have K positive whenever $m > (n + 1)$.

(3.1) THEOREM [2]. *Suppose we have an algebraic family of curves on X which cover a subvariety of X of codimension D . Suppose further that the generic member of the family is the image of a smooth algebraic curve of genus g under a locally immersive map. Then*

$$D \geq (2 - 2g)/d + m - (n + 1),$$

where d is the degree of the curves of the family (in the projective space of the hypersurface X). (This theorem is true over algebraically closed fields of arbitrary characteristics.)

In more concrete terms, this theorem says, for example, that a generic hypersurface of degree 5 or more in \mathbf{P}^3 can contain no immersed rational curve, similarly for a generic hypersurface of degree ≥ 7 in \mathbf{P}^4 , of degree ≥ 9 in \mathbf{P}^5 , etc. The proof of the theorem involves a careful comparison of normal bundles $N_{f,X} \subseteq N_{f,Y}$ (see §1), where Y is a generic hypersurface of degree m in \mathbf{P}^{n+m} and X is a linear section of Y .

4. Some interesting questions. I. Let X be a smooth projective threefold with trivial canonical bundle. Let F be a family of immersed rational curves on X . If $\dim F > 0$, is the Abel-Jacobi mapping $\phi: F \rightarrow J(X)$ always nontrivial?

I*. Let X be a smooth projective threefold with trivial canonical bundle. Let $f: \mathbf{P}^1 \rightarrow X$ be locally immersive. Is the infinitesimal Abel-Jacobi mapping

$$\phi: H^0(N_{f,X}) \rightarrow H^1(\Omega_X^2)^*$$

always injective?

(Both of these are true in all examples I know.)

If I* is true and all first-order deformations of X are unobstructed, then the generic deformation X' of X has only isolated rational curves on it, and these are all "of the first kind," that is, they have normal bundle

$$\mathcal{O}(-1) \oplus \mathcal{O}(-1). \quad (4.1)$$

Even the simplest example, V a generic hypersurface in \mathbf{P}^4 , this last statement has proved very elusive. What is known in this case is:

- (i) V admits smooth rational curves of all degrees with normal bundle (4.1),
- (ii) all rational curves of degree < 4 have normal bundle (4.1),
- (iii) V has $2875 = 23 \cdot 5^3 \cdot 1$ lines and $609,250 = 2437 \cdot 5^3 \cdot 2$ conics [6].

II. CONJECTURE. The generic quintic threefold V admits only a finite number of rational curves of each degree. Each rational curve is a smoothly embedded \mathbf{P}^1 with normal bundle (4.1). All the rational curves on V are mutually disjoint. The number of rational curves of degree d on V is

$$(\text{interesting number}) \cdot 5^3 \cdot d.$$

(The 5 comes from “quintic” and the exponent 3 comes from “threefold.” There is really very little evidence for the inclusion of the factor d in the conjecture—that part is better called a guess than a conjecture.)

III. Let $V \subseteq \mathbf{P}^4$ be a generic hypersurface of degree ≥ 6 (so that K is positive). Is every curve on V a surface section? At least, is every smooth curve C on V a “surface section to first order”—that is, does the sequence

$$0 \rightarrow N_{C,V} \rightarrow N_{C,\mathbf{P}^4} \rightarrow N_{V,\mathbf{P}^4}|_C \rightarrow 0 \quad (4.2)$$

split?

The reason one is led to ask question III is that the infinitesimal Abel-Jacobi mapping can be interpreted as an obstruction to splitting (4.2), and for generic V , the infinitesimal Abel-Jacobi mapping is presumably trivial.

BIBLIOGRAPHY

1. F. Bardelli, *On Grothendieck's generalized Hodge conjecture for a family of threefolds with geometric genus one*, Preprint, Istituto Matematico, “U. Dini,” Univ. of Florence.
2. H. Clemens, *Curves on generic hypersurfaces*, Ann. Sci. École Norm. Sup. (4) (to appear).
3. —, *The infinitesimal Abel-Jacobi mapping for hypersurfaces*, Proceedings of the Lefschetz Centennial Conference, Mexico, Dec. 1984, Contemp. Math., Amer. Math. Soc. (to appear).
4. —, *Homological equivalence, modulo algebraic equivalence, is not finitely generated*, IHES Publ. Math. **58** (1983), 19–38.
5. A. Conte and J. P. Murre, *The Hodge conjecture for fourfolds admitting a covering by rational curves*, Math. Ann. **238** (1978), 79–88.
6. S. Katz, *On the finiteness of rational curves on quintic threefolds*, Preprint, Univ. of Oklahoma.
7. S. Mori, *Projective manifolds with ample tangent bundles*, Ann. of Math. (2) **110** (1979), 593–606.
8. M. V. Nori, *The Jacobian of the generic curve of genus three*, Preprint, Tata Inst. Fund. Res., Bombay.
9. C. Schoen, *Complex multiplication cycles on elliptic modular threefolds*, Preprint, Duke Univ., Durham, N.C.

UNIVERSITY OF UTAH, SALT LAKE CITY, UTAH 84112, USA

Arithmétique des variétés rationnelles et problèmes birationnels

JEAN-LOUIS COLLIOT-THÉLÈNE

Introduction. D'un point de vue géométrique, les équations en les variables x, y, z :

$$a = bx^p + cy^q + dz^r \tag{1}_{p,q,r}$$

avec $a, b, c, d \in \mathbf{Q}$ non nuls et p, q, r entiers naturels tels que

$$1/p + 1/q + 1/r \geq 1$$

définissent des surfaces très simples, dites rationnelles: elles deviennent birationnelles au plan affine après extension finie du corps de base \mathbf{Q} . Mais classiquement, c'est seulement pour $(1)_{2,2,2}$ qu'on sait, pour toute valeur de (a, b, c, d) , traiter les problèmes diophantiens fondamentaux: Existence de solutions $(x, y, z) \in \mathbf{Q}^3$? Densité de ces solutions? Paramétrage de toutes les solutions?

Avant d'étudier une équation du type (1), il est bon de savoir mesurer sa "complexité diophantienne." On dispose pour cela d'une classification k -birationnelle des surfaces rationnelles, due à F. Enriques, B. Segre, Yu. I. Manin et V. A. Iskovskih (§1).

Le §2 décrit la méthode de la descente. Ce programme général, développé par J.-J. Sansuc et l'auteur [10, 11], et inspiré de la descente sur les courbes elliptiques, prolonge des travaux de F. Châtelet [5], P. Swinnerton-Dyer et Yu. I. Manin [35, 36]. Il réduit les problèmes diophantiens ci-dessus à deux questions fondamentales sur des variétés algébriques auxiliaires, et il a réussi pour $(1)_{2,2,3}$ et $(1)_{2,2,4}$. Ceci a permis d'établir le principe de Hasse pour essentiellement tout système de deux formes quadratiques en au moins neuf variables et a contribué à la réponse négative au problème de Zariski (§4).

La K -théorie algébrique a permis de montrer que le groupe de Chow réduit $A_0(X)$ d'une surface rationnelle X définie sur un corps de nombres est un groupe fini (§3). On a proposé une conjecture sur sa valeur, conjecture qui a été établie pour $(1)_{2,2,3}$ et $(1)_{2,2,4}$ en même temps que les résultats du §2.

Le §4 rassemble divers problèmes et résultats qui relèvent de la géométrie classique (problèmes de rationalité et de rationalité stable) mais dont les méthodes

Supported by Comité National Français des Mathématiciens.

font appel à des techniques propres au cas où le corps de base n'est pas algébriquement clos (problème de Zariski, problème de Noether, centre de l'algèbre à division générique).¹

1. La classification k -birationnelle des surfaces rationnelles. Soit k un corps parfait infini, \bar{k} une clôture algébrique, $g = \text{Gal}(\bar{k}/k)$. Une k -variété algébrique X de dimension n est dite rationnelle si $\bar{X} = X \times_k \bar{k}$ est intègre et \bar{k} -birationnelle à l'espace projectif $\mathbf{P}_{\bar{k}}^n$, et elle est dite k -rationnelle si elle est déjà k -birationnelle à l'espace \mathbf{P}_k^n . Étant donnée une k -surface rationnelle propre et lisse X et ω son fibré canonique, on appellera degré de X l'entier relatif $d = (\omega \cdot \omega) \leq 9$. On note $X(k)$ l'ensemble de ses points rationnels. On connaît deux grandes séries de surfaces rationnelles: les surfaces de Del Pezzo et les surfaces fibrées en coniques.

Une surface de Del Pezzo X est une k -surface propre et lisse sur laquelle ω^{-1} est ample [18, 36]. Sur une telle surface, $1 \leq d \leq 9$. Pour $d = 9$, on trouve les surfaces de Severi-Brauer. Les quadriques lisses de \mathbf{P}^3 ont $d = 8$. Pour $d = 4$, on trouve les intersections lisses de deux quadriques dans \mathbf{P}^4 , pour $d = 3$ les surfaces cubiques lisses dans \mathbf{P}^3 , le plongement étant défini par le système linéaire anticanonique $|\omega^{-1}|$. Pour $d = 2$, ce système fait de X un revêtement double de \mathbf{P}_k^2 ramifié le long d'une quartique lisse. Pour $d = 1$ ce système a un unique point base (donc $X(k)$ est non vide) et $|\omega^{-2}|$ fait de X un revêtement double d'un cône quadratique, ramifié en son sommet et le long d'une sextique.

On appelle surface fibrée en coniques standard une k -surface lisse X munie d'un k -morphisme propre surjectif $p: X \rightarrow C$, où la base C et la fibre générique sont des courbes propres et lisses de genre 0, et où chacune des r fibres géométriques dégénérées est l'intersection transverse de deux courbes lisses de genre zéro. Pour une telle surface, $d = 8 - r$.

La méthode d'adjonction a permis à Enriques [20], Manin [33, 34] et Iskovskih [26] de montrer que ces deux familles forment un système complet de représentants pour les classes d'équivalence k -birationnelle de k -surfaces rationnelles:

THÉOREME 1 [26]. *Toute k -surface rationnelle propre et lisse k -minimale X est soit une surface de Del Pezzo avec $\text{Pic } X \simeq \mathbf{Z}$, soit une surface fibrée en coniques standard avec $\text{Pic } X \simeq \mathbf{Z} \oplus \mathbf{Z}$.*

Ainsi, les équations $(1)_{2,3,3}$, $(1)_{3,3,3}$, resp. $(1)_{2,3,4}$, $(1)_{2,4,4}$, resp. $(1)_{2,3,5}$, $(1)_{2,3,6}$ mènent à des surfaces de Del Pezzo de degré 3, resp. 2, resp. 1 et les équations $(1)_{2,2,n}$ à des surfaces fibrées en coniques.

Le théorème 1 et une analyse cas par cas donnent le résultat suivant, conjecturé par Manin [32] (et qu'on aimerait étendre au cas où $cdk \leq 1$):

PROPOSITION 2. *Si k est un corps C_1 , toute k -surface rationnelle propre et lisse possède un point k -rationnel.*

¹On trouvera dans le texte récent de Manin et Tsfasman [37] de nombreux compléments au présent rapport.

Il est utile de savoir quand des surfaces d'un certain type sont k -birationnelles à celles d'un autre type (en particulier si elles sont k -rationnelles). Pour $d \geq 5$, la réponse est simple: toute k -surface rationnelle propre et lisse de degré ≥ 5 est k -rationnelle dès que $X(k)$ est non vide. De plus, si k est un corps de nombres, le principe de Hasse vaut pour de telles surfaces (Manin [33]). Pour $d \leq 4$ commencent les problèmes arithmétiques sérieux. En utilisant la technique des *systèmes linéaires à points bases*, et suivant B. Segre [49], Manin [34] et Iskovskih [23, 24, 25, 27] ont résolu les problèmes de classification k -birationnelle. Leurs principaux résultats peuvent s'énoncer ainsi [37, 50]:

THÉOREME 3. *Soient X et Y deux k -surfaces rationnelles k -minimales avec X de degré ≤ 4 . Si $f: X \rightarrow Y$ est une k -application birationnelle, alors f peut s'écrire comme un composé d'applications k -birationnelles*

$$X = X_0 \xrightarrow{f_1} X_1 \xrightarrow{f_2} \dots \xrightarrow{f_n} X_n = Y$$

de telle sorte que pour chaque i , ou bien X_{i-1} est k -isomorphe à X_i , ou bien X_{i-1} et X_i sont des fibrations en coniques standard et f_i est une transformation élémentaire en un point fermé (non situé sur une fibre dégénérée)

Joint à une description des surfaces qui peuvent à la fois être des fibrations en coniques relativement minimales et des surfaces de Del Pezzo [26], cet énoncé implique (cf. Skorobogatov [50]):

COROLLAIRE 4. *Conservons les hypothèses du Théorème 3. La surface X n'est pas k -rationnelle. Les degrés de X et Y coïncident. Les \mathfrak{g} -modules $\text{Pic } \bar{X}$ et $\text{Pic } \bar{Y}$ sont isomorphes, ainsi que les groupes $\text{Pic } X$ et $\text{Pic } Y$. Si X est une surface de Del Pezzo de degré $d \leq 3$ avec $\text{Pic } X \simeq \mathbf{Z}$ (ce qui est le cas si $d = 3$), X n'est pas k -birationnelle à une surface fibrée en coniques. Si X est une surface de Del Pezzo de degré 4 avec $\text{Pic } X \simeq \mathbf{Z}$, elle n'est pas k -birationnelle à une surface fibrée en coniques de même degré.*

Pour X/C une surface fibrée en coniques relativement minimale de degré $d \leq 3$, on a mieux (Iskovskih): la base et la fibre générique de la fibration ne dépendent que de la classe d'équivalence k -birationnelle de la surface X . Pour $d = 4$, on doit à Manin [33], Kunyavskiĭ, Skorobogatov, Tsfasman [30, 31] une classification combinatoire fine suivant l'action de \mathfrak{g} sur $\text{Pic } \bar{X}$.

On dit qu'une k -variété X est k -unirationnelle s'il existe une application k -rationnelle dominante d'un espace projectif \mathbf{P}_k^r sur X . Les points k -rationnels de X sont alors Zariski-denses. Sauf en ce qui concerne les surfaces fibrées en coniques de degré 1 ou 2, les résultats ci-après sont classiques ou dus à Manin [34]:

THÉOREME 5. *Soit X une k -surface rationnelle, propre et lisse, de degré d , possédant un point k -rationnel. Si $d \geq 5$, X est k -rationnelle, et si $d = 4$ ou 3, elle est k -unirationnelle. Si $d = 2$ ou si X est une surface fibrée en coniques standard avec $d = 1$, et si de plus X possède des points k -rationnels en dehors de certaines courbes, alors X est k -unirationnelle.*

Les problèmes importants de la k -unirationalité et de la densité des points rationnels restent ouverts tant pour les surfaces fibrées en coniques de degré $d \leq 0$ (e.g. $(1)_{2,2,n}$ pour $n \geq 8$) que pour les surfaces de Del Pezzo de degré 1 (qui possèdent toujours un k -point!) (ainsi on ne sait pas si les équations originelles $(1)_{2,3,5}$ et $(1)_{2,3,6}$, qui sont des ouverts de surfaces de Del Pezzo de degré 1, ont toujours une solution). On ne sait pas non plus si $(1)_{2,3,4}$, qui est un ouvert d'une surface de Del Pezzo de degré 2 possédant un k -point, a toujours une solution rationnelle.

2. La descente: programme, résultats, perspectives.

2.1. *Le module galoisien* $\text{Pic } \overline{X}$. Le \mathfrak{g} -module $\text{Pic } \overline{X}$ est apparu à plusieurs titres dans les travaux de Manin et de Voskresenskiï [32, 36, 53, 56] sur les variétés rationnelles. Son premier rôle est birationnel:

THÉORÈME 6. *Soient X et Y des k -variétés propres, lisses, géométriquement intègres, rationnelles. Alors:*

- (i) *Les \mathfrak{g} -modules $\text{Pic } \overline{X}$ et $\text{Pic } \overline{Y}$ sont des \mathfrak{g} -modules \mathbf{Z} -libres de type fini.*
- (ii) *S'il existe des entiers n et m tels que $X \times_k \mathbf{P}_k^n$ soit k -birationnelle à $Y \times_k \mathbf{P}_k^m$, alors $\text{Pic } \overline{X}$ et $\text{Pic } \overline{Y}$ sont isomorphes à addition près de \mathfrak{g} -modules de permutation. En particulier, si X est stablement k -rationnelle, $\text{Pic } \overline{X}$ est stablement de permutation.*
- (iii) (Voskresenskiï) *Si X et Y sont des compactifications lisses de k -tores algébriques, la réciproque de (ii) vaut.*
- (iv) *Sous les hypothèses de (ii), les groupes $H^1(\mathfrak{g}, \text{Pic } \overline{X})$ et $H^1(\mathfrak{g}, \text{Pic } \overline{Y})$ coïncident et $H^1(\mathfrak{g}, \text{Pic } \overline{X}) = 0$ si X est stablement k -rationnelle.*

Notons $\text{Br } X = H_{\text{ét}}^2(X, \mathbf{G}_m)$ le groupe de Brauer-Grothendieck. Pour X comme ci-dessus, on dispose d'une injection naturelle $\text{Br } X / \text{Br } k \hookrightarrow H^1(\mathfrak{g}, \text{Pic } \overline{X})$ qui est un isomorphisme si $X(k)$ est non vide ou si k est un corps de nombres. C'est par le biais de ce lien à $\text{Br } X$ que l'on voit apparaître le second rôle de $\text{Pic } \overline{X}$, rôle dégagé par Manin [35, 36], et qui est de donner de l'information sur l'ensemble $X(k)$ des points rationnels de X , tant sur leur existence que sur la répartition de $X(k)$ en classes pour la R -équivalence.

Soit X une k -variété propre et lisse sur un corps de nombres k . Si pour tout $(P_v) \in \prod_v X(k_v) \neq \emptyset$, il existe $\mathcal{A} \in \text{Br } X$ avec $\sum_v i_v(\mathcal{A}(P_v)) \neq 0$, où $\mathcal{A}(P_v) \in \text{Br } k_v$ est la fibre de \mathcal{A} en $P_v \in X(k_v)$ et où $i_v: \text{Br } k_v \rightarrow \mathbf{Q}/\mathbf{Z}$ est l'invariant local en v , alors $X(k) = \emptyset$: c'est là l'obstruction de Manin [35] au principe de Hasse pour X , obstruction qui rend compte de multiples contre-exemples à ce "principe."

On dit que l'obstruction de Manin au principe de Hasse est *la seule* pour une classe de k -variétés propres et lisses si pour toute X dans cette classe avec $X(k) = \emptyset$ et $\prod_v X(k_v) \neq \emptyset$, la condition ci-dessus est satisfaite. Sansuc [47] a ainsi établi que cette obstruction est la seule pour la classe \mathcal{C} des (compactifications lisses des) espaces principaux homogènes sous un groupe linéaire connexe sans facteur de type E_8 .

Manin [36] a aussi interprété et partiellement généralisé les exemples de Châtelet [5] au moyen de l'équivalence de Brauer sur $X(k)$ (obtenue par accouplement avec $\text{Br } X$). Mais au moins pour les variétés rationnelles, ses résultats sont couverts par la méthode de la descente décrite ci-dessous. Rappelons la définition qu'il a introduite dans ce contexte: deux k -points P et Q d'une k -variété Y sont R -équivalents s'il existe une suite finie $P = P_0, \dots, P_i, P_{i+1}, \dots, P_n = Q$ de points de $Y(k)$ telle que pour tout i , P_i et P_{i+1} appartiennent à $\varphi_i(U_i(k))$, pour φ_i un k -morphisme d'un ouvert U_i de \mathbf{P}_k^1 vers Y .

2.2. *Le programme de la descente* [10, 11]. Soit S un k -tore, \hat{S} son groupe des caractères, qui est un \mathfrak{g} -module \mathbf{Z} -libre de type fini. Si X est une k -variété, les toiseurs sur X sous S sont classifiés par le groupe de cohomologie étale $H^1(X, S)$, et l'homomorphisme naturel $\chi: H^1(X, S) \rightarrow \text{Hom}_{\mathfrak{g}}(\hat{S}, \text{Pic } \bar{X})$ associe à la classe de tout toiseur son *type*. Supposons désormais que X est une k -variété propre et lisse rationnelle. On peut considérer le k -tore S_0 tel que $\hat{S}_0 = \text{Pic } \bar{X}$ puisque ce dernier \mathfrak{g} -module est alors \mathbf{Z} -libre de type fini. On appelle *toiseur universel* un toiseur sur X sous S_0 dont le type est l'identité de \hat{S}_0 . Tout toiseur τ sur X définit par spécialisation une application $\theta_\tau: X(k) \rightarrow H^1(k, S)$ qui associe à un k -point P la classe de la fibre $\tau(P)$ de τ en P . Cette classe est triviale si et seulement s'il existe un k -point sur la k -variété τ qui s'envoie sur P par la projection structurale $p_\tau: \tau \rightarrow X$. Tordre un toiseur τ par l'opposé d'un élément $\alpha \in H^1(k, S)$ définit un toiseur τ^α de même type. Étant donné un type $\lambda \in \text{Hom}_{\mathfrak{g}}(\hat{S}, \text{Pic } \bar{X})$, on a la décomposition (union disjointe):

$$X(k) = \bigcup_{\tau \text{ de type } \lambda} p_\tau(\tau(k)) = \bigcup_{\alpha \in H^1(k, S)} p_{\tau^\alpha}(\tau^\alpha(k)),$$

où l'on peut se contenter d'écrire la réunion sur les α dans l'image de θ_τ .

THÉORÈME 7. *Cette décomposition est plus grossière que la décomposition en classes pour la R -équivalence. La décomposition associée aux toiseurs universels est plus fine que celle associée à tout autre type, et elle est plus fine que l'équivalence de Brauer. Elle est finie si k est un corps de nombres. Sous cette même hypothèse, si $X(k_v) \neq \emptyset$ pour toute place v de k et s'il n'y a pas d'obstruction de Manin au principe de Hasse pour X , pour tout type λ il existe un toiseur τ de ce type avec $\tau(k_v) \neq \emptyset$ pour toute place v .*

La démonstration de ce théorème utilise en particulier les théorèmes de dualité de Tate-Nakayama en théorie du corps de classes.

L'étude des points rationnels sur X est ainsi ramenée à l'étude des points rationnels sur certaines variétés auxiliaires, en principe plus simples:

THÉORÈME 8. *Sur les (k -compactifications lisses des) toiseurs universels (lesquels sont des k -variétés rationnelles) les obstructions tant à la k -rationalité (Théorème 6) qu'au principe de Hasse (obstruction de Manin) liées au module de Picard disparaissent.*

Etant données une classe C de k -variétés propres et lisses rationnelles, on considère alors les deux hypothèses fondamentales :

(H1) *Les toreseurs universels sur X dans C qui possèdent un point k -rationnel sont des k -variétés k -rationnelles.*

(H2) *Si k est un corps de nombres, le principe de Hasse vaut pour les toreseurs universels sur X dans C .*

Si ces deux hypothèses valent, le programme de la descente réussit parfaitement : l'arithmétique des variétés de la classe C est alors essentiellement connue, au moins d'un point de vue qualitatif [11]. L'obstruction de Manin au principe de Hasse sur une variété X de C est alors la seule obstruction possible, et $X(k)$ se découpe en un nombre fini de classes pour la R -équivalence, chacune paramétrée par les k -points d'une k -variété k -rationnelle.

Pour essayer d'établir (H1) ou (H2), il est important d'avoir des équations concrètes des toreseurs universels : on dispose pour cela d'"équations locales" qui permettent de les représenter comme des produits fibrés convenables.

2.3. Les résultats.

2.3.1. *Les tores* [12, 11, 15, 56, 52]. Si X est une k -compactification lisse d'un k -tore T , on dispose de la suite exacte de \mathfrak{g} -modules \mathbf{Z} -libres de type fini de Voskresenskii [53, 56] :

$$0 \rightarrow \hat{T} \rightarrow \mathrm{Div}_{\infty} \bar{X} \rightarrow \mathrm{Pic} \bar{X} \rightarrow 0,$$

et de la suite duale de k -tores : $1 \rightarrow S \rightarrow P \rightarrow T \rightarrow 1$.

La description locale des toreseurs universels révèle que le toseur sur T sous S défini par cette suite est la restriction à l'ouvert $T \subset X$ du toseur universel sur X trivial en $0 \in T(k)$. Comme P est un tore quasi-trivial (dual du \mathfrak{g} -module de permutation défini par les diviseurs à support hors de T), on conclut que T est une k -variété k -rationnelle. Un argument similaire vaut pour les autres toreseurs universels, qui ici possèdent tous un k -point. Le programme de la descente réussit ici parfaitement, et mène à l'isomorphisme $T(k)/R \simeq H^1(k, S)$ qui permet de calculer effectivement la R -équivalence (finie si k est un corps de nombres) sur $T(k)$.

2.3.2. *Les surfaces de Châtelet généralisées* [16, 8]. Il s'agit des modèles propres et lisses des surfaces d'équation affine (2), généralisant (1)_{2,2,3} et (1)_{2,2,4} :

$$y^2 - az^2 = P(x) \quad (a \in k^*), \quad (2)$$

où $P(x)$ est un polynôme séparable de degré 3 ou 4. On trouvera dans [17] une description de nombreux modèles naturels, singuliers ou non, de ces surfaces. Du point de vue de la classification du §1, ce sont les surfaces rationnelles non triviales les plus simples (fibrations en coniques au-dessus de \mathbf{P}_k^1 , de degré 4 et possédant une section sur une extension quadratique de k).

Dans [16], en collaboration avec Swinnerton-Dyer, on établit (H1) et (H2) pour ces surfaces. Pour cela, on montre en utilisant les équations locales des toreseurs universels que ce sont essentiellement des intersections de deux quadriques dans \mathbf{P}_k^7 , intersections qui possèdent une paire de sous-espaces linéaires

gauches conjugués. Un argument géométrique facile donne alors (H1). Pour (H2), on établit le principe de Hasse sur une intersection de deux quadriques dans \mathbf{P}_k^n ($n \geq 4$) possédant une paire de droites conjuguées, et n'étant pas d'un type exceptionnel si $n = 5$, en se réduisant au cas connu $n = 4$ par *sections hyperplanes* définies sur k et contenant les deux droites. Ce processus, où l'on doit conserver l'hypothèse d'existence de solutions locales partout, est un peu délicat (il y a des contre-exemples pour $n = 5$). On le mène à bien par une analyse très précise des équations.

Une méthode très différente, développée avec D. Coray [8], permet d'établir le principe de Hasse pour certaines intersections de deux quadriques et avait donné les premiers résultats arithmétiques sur les équations (2). Cette méthode est fondée sur la théorie des formes quadratiques sur $k(t)$ et sur un théorème de Amer et Brumer.

L'arithmétique des surfaces (2) est ainsi parfaitement comprise. Ainsi:

THÉOREME 9. *Si $P(x)$ est un polynôme irréductible du 4^{ème} degré, l'équation (2) satisfait le principe de Hasse.*

En utilisant ce nouveau cas du principe de Hasse comme constituant élémentaire dans la technique des *sections hyperplanes*, ainsi que le théorème d'irréductibilité de Hilbert, on montre:

THÉOREME 10. *Deux formes quadratiques en au moins neuf variables sur un corps de nombres k totalement imaginaire ont un zéro commun non trivial.*

Si le corps de nombres k est réel, les obstructions de nature réelle à l'existence d'un zéro commun sont, à des détails techniques près, les seules.

Coray et Tsfasman [17] ont établi que l'arithmétique des surfaces cubiques singulières non coniques et des intersections singulières non coniques de deux quadriques dans \mathbf{P}_k^4 se ramène à celle des surfaces rationnelles lisses de degré $d \geq 5$ ou à celle des surfaces de Châtelet généralisées: ainsi l'arithmétique de telles surfaces est aussi parfaitement comprise.

2.4. Les perspectives. L'hypothèse (H1) ne vaut pas toujours pour $\dim X > 2$, mais la question est ouverte en dimension deux, où le \mathfrak{g} -module $\text{Pic } \overline{X}$ semble fournir un grand contrôle de X (§3). Notons néanmoins que (H1) implique la k -unirationalité des surfaces X avec $X(k) \neq \emptyset$! Faute de pouvoir établir d'emblée (H1) ou (H2) pour toutes les surfaces rationnelles, la première tâche est d'étudier ces questions sur les classes les plus simples de surfaces rationnelles.

2.4.1. Sur une fibration en coniques standard quelconque X/\mathbf{P}_k^1 de degré $8 - r$, l'étude des toseurs universels se ramène à celle de certaines intersections de $r - 2$ quadriques dans \mathbf{P}_k^{2r-1} [10]. Pour $r = 4$, ceci a déjà permis d'obtenir d'autres résultats [41], en particulier la finitude de la R -équivalence [10]. Pour pouvoir aller plus loin, il faudrait connaître le principe de Hasse "lisse" pour les intersections non coniques de deux quadriques dans \mathbf{P}_k^7 . Le cas $r = 5$, qui correspond aux surfaces cubiques lisses avec une droite définie sur k , mène à des intersections de 3 quadriques dans \mathbf{P}_k^9 .

2.4.2. En utilisant comme constituant élémentaire les surfaces cubiques avec exactement trois points singuliers conjugués (Skolem, cf. [17]) on peut développer la méthode des sections hyperplanes comme dans [16]: on peut ainsi établir le principe de Hasse lisse pour la plupart des hypersurfaces cubiques possédant trois points singuliers conjugués, et développer sur les hypersurfaces d'équation affine $N_{K/k}(x + \omega y + \omega^2 z) = P(t)$, où $K = k(\omega)$ est une extension cubique de k et $P(t)$ est un polynôme de degré 2 ou 3, une théorie analogue à celle des surfaces de Châtelet généralisées (travail en préparation avec P. Salberger).

2.4.3. Il est souhaitable de calculer explicitement l'obstruction de Manin dans des exemples concrets. C'est facile dans le cas $(1)_{2,2,n}$, et dans ce cas on a établi, pour $k = \mathbf{Q}$, un lien entre (H2) et l'hypothèse de Schinzel sur les premiers représentés par des polynômes [14]. Dans le cas $(1)_{3,3,3}$ (surfaces cubiques diagonales sur \mathbf{Q}), c'est moins facile, mais cela a été mené à bien avec D. Kanevsky dans [9]. Ceci a permis des calculs numériques (M. Vallino) qui ont conforté l'hypothèse que l'obstruction de Manin au principe de Hasse pour $(1)_{3,3,3}$ est la seule. Il en serait ainsi si certaines intersections très particulières de deux hypersurfaces cubiques dans \mathbf{P}_k^8 satisfaisaient le principe de Hasse [9].

2.4.4. Si l'obstruction de Manin au principe de Hasse pour les intersections lisses de deux quadriques dans \mathbf{P}_k^4 était la seule, on en déduirait par la méthode des sections hyperplanes que le principe de Hasse vaut pour les intersections lisses de deux quadriques dans \mathbf{P}_k^n pour $n \geq 5$. Pour d'autres conséquences du même type, voir l'exposé de Kanevsky [28].

3. Groupe de Chow et K -théorie. On note $\mathrm{CH}_0(X)$ le groupe de Chow des 0-cycles (combinaisons linéaires à coefficients entiers de points fermés) modulo l'équivalence rationnelle sur une k -variété X et, si X est propre, on note $A_0(X)$ le sous-groupe des classes de 0-cycles de degré 0.

3.1. *Surfaces rationnelles.* S. Bloch [2] montra comment la cohomologie galoisienne du complexe des sections sur \overline{X} de la résolution de Quillen du faisceau \mathcal{K}_2 donne naissance, lorsque X est une k -surface rationnelle propre et lisse, à une suite exacte

$$\cdots \rightarrow H^1(\mathfrak{g}, K_2 \overline{k}(X)) \rightarrow A_0(X) \xrightarrow{\Phi} H^1(k, S_0) \rightarrow \cdots \quad (3)$$

où S_0 est le k -tore dual de $\mathrm{Pic} \overline{X}$ (l'homomorphisme caractéristique Φ n'est pas sans rapport avec les toreseurs universels, voir [13]). Des arguments de bonne réduction montrent alors que si k est un corps de nombres, Φ a une image finie, et des résultats de Merkur'ev/Suslin [38] impliquent [6] que le groupe $H^1(\mathfrak{g}, K_2 \overline{k}(X))$ est nul pour toute k -variété X possédant un k -point lisse, et que Φ est toujours injective si k est un corps de nombres: *le groupe $A_0(X)$ est donc fini pour toute surface rationnelle définie sur un corps de nombres* (résultat obtenu d'abord par Bloch [2] pour X fibrée en coniques sur \mathbf{P}_k^1 , via des techniques de formes quadratiques). Un résultat de finitude similaire vaut si k est un corps local, et le groupe $A_0(X)$ est alors nul si X a bonne réduction.

Les dualités locales et globales de Tate-Nakayama permettent d'intégrer la flèche diagonale $H^1(k, S_0) \rightarrow \coprod_v H^1(k_v, S_0)$, de noyau $\mathrm{III}^1(k, S_0)$ (v parcourt

les places de k), dans une longue suite exacte. Sansuc et moi-même [13, 48] conjecturons que, via Φ , cette suite induit une suite exacte de groupes finis:

$$(H3) \quad 0 \rightarrow \text{III}^1(k, S_0) \rightarrow A_0(X) \xrightarrow{\iota} \coprod_v A_0(X_v) \rightarrow \text{Hom}(H^1(\mathfrak{g}, \text{Pic } \overline{X}), \mathbf{Q}/\mathbf{Z}).$$

Cette conjecture assurerait que le groupe global $A_0(X)$ est gros dès qu'il en est ainsi de la somme des groupes locaux (les termes extrêmes étant peu affectés par l'arithmétique de la surface). Si $X(k)$ est non vide, (H2) (§2) donne pour $\text{Ker } \iota$ la valeur prédite par (H3). Dans [16], on établit en fait toute la conjecture (H3) pour les surfaces de Châtelet généralisées X avec $X(k) \neq \emptyset$, en utilisant en outre un résultat d'effectivité sur les 0-cycles de degré ≥ 1 ([7], voir aussi [39] pour une démonstration via (4) ci-dessous et le théorème de Riemann-Roch de Witt).

Il serait intéressant de calculer les groupes locaux $A_0(X_v)$ en fonction de la mauvaise réduction de X_v (travaux en cours de Bloch, Coombes, Muder).

3.2. Généralisations.

3.2.1. Soit Y un schéma de Dedekind intègre de corps des fonctions K , à corps résiduels parfaits, soit A une algèbre simple centrale sur K , d'indice n et soit Λ un faisceau d'ordres héréditaires dans A . Le groupe des classes $\text{Cl}(\Lambda)$ est le sous-groupe de rang 0 du groupe de Grothendieck $K_0^{\text{lf}}(\Lambda)$ des Λ -modules \mathcal{O}_Y -cohérents localement libres. Lorsque $n = 2$, on associe à Λ une fibration en coniques standard X_Λ sur Y , et Salberger [39] a établi un isomorphisme

$$\text{CH}_0(X_\Lambda) \simeq \text{Cl}(\Lambda) \quad (4)$$

qui doit admettre des généralisations pour $n > 2$. Si k est un corps de nombres, $n = 2$ et Y est la droite affine \mathbf{A}_k^1 , le résultat de finitude indiqué plus haut implique alors que $\text{Cl}(\Lambda)$ est fini. Dans [40], en ayant recours à la première définition de la K -théorie pour définir un nouvel homomorphisme caractéristique ainsi qu'à un résultat de Merkur'ev/Suslin, Salberger montre que ce résultat de finitude s'étend au cas où $Y = \mathbf{A}_k^1$ et où $n > 2$ est sans facteur carré.

3.2.2. Lorsque k est un corps de nombres, une conjonction de la méthode de Bloch et d'un résultat récent de K. Kato sur la cohomologie étale des courbes a permis à M. Gros [21] et à T. Ôkochi de montrer que le groupe $A_0(X)$ est un groupe de type fini quand X est une surface fibrée en coniques au-dessus d'une courbe de genre quelconque. La méthode devrait permettre d'étendre le résultat de génération finie de 3.2.1 au cas où la courbe de base Y est une courbe lisse affine quelconque.

3.2.3. Le groupe $A_0(X)$ est fini si X est projective et lisse sur un corps fini (Kato, Saito) et il est nul si X est rationnelle (Bloch). Pour certaines surfaces X sur k p -adique, on dispose de résultats de finitude pour $A_0(X)_{\text{tors}}$.

4. Questions de rationalité en géométrie algébrique.

4.1. *Le problème de Zariski.* On dit qu'une k -variété intègre X de dimension d est stablement k -rationnelle s'il existe un entier n que $X \times_k \mathbf{P}_k^n$ soit k -birationnel à \mathbf{P}_k^{n+d} . C'était une question de Zariski (1949) de savoir si toute telle k -variété est déjà k -rationnelle. Il en est ainsi si $d = 1$ (Lüroth) ou si k est algébriquement

clos et $d = 2$ (Castelnuovo). Dans l'article commun [1], nous montrons que la réponse est négative si $d = 2$ et k non algébriquement clos convenable, et si $d = 3$ et $k = \mathbf{C}$. Pour $d = 2$, on considère un modèle propre et lisse X de la surface de Châtelet (2) avec $a \in k^*$ non carré et $P(x)$ polynôme irréductible du 3^{ème} degré en x , de discriminant égal à a . Le \mathfrak{g} -module $\text{Pic } \bar{X}$ est alors stablement de permutation, et comme on connaît (H1) (voir §2.3.2), on conclut que X est stablement k -rationnelle; le Corollaire 4 montre par ailleurs que la surface X n'est pas k -rationnelle. On obtient ainsi des exemples avec $k = \mathbf{Q}$ et $k = \mathbf{C}(t)$, auquel cas la variété de dimension 3 sur \mathbf{C} définie par (2) est stablement \mathbf{C} -rationnelle. Dans ce dernier cas, la méthode de la jacobienne intermédiaire (Clemens/Griffiths, Mumford, Beauville) et un choix convenable du polynôme $P(x) \in \mathbf{C}(t)[x]$ montrent qu'un modèle projectif et lisse Y de la variété de dimension 3 définie par (2) n'est pas \mathbf{C} -rationnelle.

4.2. *Le problème de Noether: corps d'invariants d'un groupe fini.* Soit k un corps, G un groupe fini, $k(G)$ le corps des fractions de l'algèbre symétrique de la représentation régulière de G sur k , et soit $L = k(G)^G$ le corps des invariants. Le corps L est-il transcendant pur sur k ? Il en est ainsi lorsque k est algébriquement clos et G abélien (Fischer, 1912). Si G est abélien et k non algébriquement clos, la réponse peut être négative (Swan [51, 52], Voskresenskiĭ [54, 55, 56]). On dispose de critères très complets de rationalité dans ce cas (Endo/Miyata, Lenstra, Voskresenskiĭ) car, comme l'a remarqué Voskresenskiĭ, L n'est autre que le corps des fonctions d'un certain k -tore T_G , auquel on peut appliquer le Théorème 6. Le plus petit contre-exemple est $k = \mathbf{Q}$, $G = \mathbf{Z}/8$ qui, comme l'a remarqué Saltman [42, 52], peut être ramené au fameux contre-exemple de Wang au théorème de Grunwald: le \mathbf{Q} -tore $T_{\mathbf{Z}/8}$ est loin d'être \mathbf{Q} -rationnel, car il ne satisfait pas l'approximation faible. Saltman [42, 43] a exploré plus avant les liens qui relient la rationalité et les propriétés d'approximation faible, ou de relèvement (voir aussi [15]).

Saltman [44, 46] a récemment donné une réponse négative au problème d'Emmy Noether lorsque k est algébriquement clos. Ses exemples reposent sur le groupe de Brauer "non-ramifié" $\text{Br}_{\text{nr}} L$ d'un corps de fonctions L ($= \text{Br } X$ pour X un modèle projectif et lisse du corps de fonctions L —i.e. l'invariant utilisé par Artin/Mumford dans leur réponse négative au problème de Lüroth). L'exemple de [46] est un sous-produit de l'étude des groupes $\text{Br}_{\text{nr}} \mathbf{C}(M)^G$ pour $\mathbf{C}(M)$ le corps des fractions de l'algèbre de groupes sur \mathbf{C} d'un G -module fidèle \mathbf{Z} -libre de type fini M . On aimerait connaître la valeur exacte de ces groupes $\text{Br}_{\text{nr}} \mathbf{C}(M)^G$.

4.3. *Corps d'invariants d'un groupe linéaire connexe.* Étant donné un groupe linéaire connexe G et une action linéaire presque libre de G sur un vectoriel complexe V , on peut se demander si le "quotient" V/G est une variété rationnelle. Ce problème apparaît naturellement dans l'étude de nombreux espaces de modules (Katsylo, Bogomolov, Dolgachev [3, 29, 19]). La rationalité *stable* de V/G ne dépend en fait pas de la représentation presque libre V (voir [19]). Elle est connue dans plusieurs cas si G est semisimple et simplement connexe (Bogomolov [4]). Quand G n'est pas simplement connexe, la situation est plus délicate: un

cas célèbre sans réponse est celui de l'action de PGL_n sur les paires de matrices (n, n) , définie par $g(A, B) = (gAg^{-1}, gBg^{-1})$ (problème du centre de l'algèbre à division générique). En utilisant une réduction due à Procesi et Formanek on a pu montrer [45, 15] que le groupe de Brauer non-ramifié du quotient est trivial si bien que la méthode d'Artin/Mumford ne saurait donner la réponse.

Il serait temps d'essayer de calculer effectivement, tant sur les exemples de Saltman que sur le centre de l'algèbre à division générique, les invariants birationnels cohomologiques supérieurs proposés par Grothendieck [22, §9].

Je remercie Jean-Jacques Sansuc pour sa constante collaboration.

BIBLIOGRAPHIE

1. A. Beauville, J.-L. Colliot-Thélène, J.-J. Sansuc et Sir Peter Swinnerton-Dyer, *Variétés stablement rationnelles non rationnelles*, Ann. of Math. **121** (1985), 283–318.
2. S. Bloch, *On the Chow groups of certain rational surfaces*, Ann. Sci. École Norm. Sup. (4) **14** (1981), 41–59.
3. F. A. Bogomolov and P. I. Katzylo, *Rationalité de quelques variétés quotients*, Mat. Sb. **126** (1985), 584–589=Math. USSR-Sb. **54** (1986), 571–576.
4. F. A. Bogomolov, *Rationalité stable des espaces-quotients des groupes simplement connexes*, Mat. Sb. **130** (1986), 3–17.
5. F. Châtelet, *Points rationnels sur certaines courbes et surfaces cubiques*, Enseign. Math. (2) **5** (1959), 153–170.
6. J.-L. Colliot-Thélène, *Hilbert's theorem 90 for K_2 , with application to the Chow groups of rational surfaces*, Invent. Math. **71** (1983), 1–20.
7. J.-L. Colliot-Thélène et D. Coray, *L'équivalence rationnelle sur les points fermés des surfaces rationnelles fibrées en coniques*, Compositio Math. **39** (1979), 301–332.
8. J.-L. Colliot-Thélène, D. Coray et J.-J. Sansuc, *Descente et principe de Hasse pour certaines variétés rationnelles*, J. Reine Angew. Math. **320** (1980), 150–191.
9. J.-L. Colliot-Thélène, D. Kanevsky et J.-J. Sansuc, *Arithmétique des surfaces cubiques diagonales*, Lecture Notes in Math. (G. Wüstholz, ed.), Springer-Verlag (à paraître).
10. J.-L. Colliot-Thélène et J.-J. Sansuc, C. R. Acad. Sci. Paris Sér. A-B **282** (1976), A1113–A1116; **284** (1977), A967–A970, A1215–A1218; **303** (1986), I303–I306.
11. —, *La descente sur les variétés rationnelles*, Journées de Géométrie Algébrique d'Angers (A. Beauville, éd.), Sijthoff and Noordhoff, Alphen aan den Rijn, 1980, pp. 223–237; *La descente sur les variétés rationnelles. II*, Duke Math. J. **54** (1987) (à paraître).
12. —, *La R-équivalence sur les tores*, Ann. Sci. École Norm. Sup. (4) **10** (1977), 175–229.
13. —, *On the Chow groups of certain rational surfaces: a sequel to a paper of S. Bloch*, Duke Math. J. **48** (1981), 421–447.
14. —, *Sur le principe de Hasse et l'approximation faible, et sur une hypothèse de Schinzel*, Acta Arith. **41** (1982), 33–53.
15. —, *Principal homogeneous spaces under flasque tori, applications*, J. Algebra **106** (1987), 148–205.
16. J.-L. Colliot-Thélène, J.-J. Sansuc and Sir Peter Swinnerton-Dyer, *Intersections of two quadrics and Châtelet surfaces*, J. Reine Angew. Math. I, **373** (1987), 37–107; II, **374** (1987), 72–168 (cf. C. R. Acad. Sci. Paris Sér. I Math. **298** (1984), 377–380).
17. D. F. Coray and M. A. Tsfasman, *Arithmetic on singular Del Pezzo surfaces*, Proc. London Math. Soc. (to appear).
18. M. Demazure, *Surfaces de Del Pezzo. II, III, IV, V*, Séminaire sur les Singularités des Surfaces, Lecture Notes in Math., vol. 777, Springer-Verlag, 1980, pp. 23–69.
19. I. Dolgachev, *Rationality of fields of invariants*, Proceedings of the 1985 Bowdoin Conference on Algebraic Geometry.

20. F. Enriques, *Sulle irrazionalità da cui può farsi dipendere la risoluzione d'un' equazione algebrica $f(x, y, z) = 0$ con funzioni razionali di due parametri*, Math. Ann. **49** (1897), 1–23.
21. M. Gros, *0-cycles de degré 0 sur les surfaces fibrées en coniques*, J. Reine Angew. Math. **373** (1987), 166–184.
22. A. Grothendieck, *Le groupe de Brauer, III : Exemples et compléments*, Dix Exposés sur la Cohomologie des Schémas, North-Holland, Amsterdam; Masson, Paris, 1968.
23. V. A. Iskovskih, *Surfaces rationnelles avec un pinceau de courbes rationnelles*, Mat. Sb. **74** (1967), 608–638=Math. USSR-Sb. **3** (1967), 563–587.
24. —, *Surfaces rationnelles avec un pinceau de courbes rationnelles dont le carré de la classe canonique est positif*, Mat. Sb. **83** (1970), 90–119=Math. USSR-Sb. **12** (1970), 91–117.
25. —, *Propriétés birationnelles des surfaces de degré 4 dans P_k^4* , Mat. Sb. **88** (1972), 31–37=Math. USSR-Sb. **17** (1972), 575–577.
26. —, *Modèles minimaux des surfaces rationnelles sur les corps quelconques*, Izv. Akad. Nauk SSSR Ser. Mat. **43** (1979), 19–43=Math. USSR-Izv. **14** (1980), 17–39.
27. —, *Générateurs et relations dans le groupe des automorphismes birationnels de deux classes de surfaces rationnelles*, Trudy Mat. Inst. Steklov. **165** (1984), 67–78=Proc. Steklov Inst. Math. **3** (1985), 73–84.
28. D. Kanevsky, *Applications of the conjecture on the Manin obstruction to various diophantine problems*, Journées Arithmétiques de Besançon (juin 1985), Astérisque **147-148** (1987), 307–314.
29. P. I. Katsylo, *Rationalité des espaces de modules de courbes hyperelliptiques*, Izv. Akad. Nauk SSSR Ser. Mat. **48** (1984), 705–710=Math. USSR-Izv. **25** (1985), 45–50.
30. B. È. Kunyavskii and M. A. Tsfasman, *Zéro-cycles sur les surfaces rationnelles et tores de Néron-Severi*, Izv. Akad. Nauk SSSR Ser. Mat. **48** (1984), 631–654=Math. USSR-Izv. **24** (1985), 583–603.
31. B. È. Kunyavskii, A. N. Skorobogatov and M. A. Tsfasman, *Combinatoire et géométrie des surfaces de Del Pezzo de degré 4*, Uspekhi Mat. Nauk **40** (1985), 145–146=Russian Math. Surveys **40** (1985), 131–132.
32. Yu. I. Manin, *Rational surfaces and Galois cohomology*, Actes du Congrès Intern. Math. (Moscou, 1966), Izdat. "Mir", Moscou, 1968, pp. 495–509.
33. —, *Surfaces rationnelles sur les corps parfaits*, Inst. Hautes Études Sci. Publ. Math. **30** (1966), 55–113=Transl. Amer. Math. Soc. (2) **84** (1969), 137–186.
34. —, *Surfaces rationnelles sur les corps parfaits. II*, Mat. Sb. **72** (1967), 161–192=Math. USSR-Sb. **1** (1967), 141–168.
35. —, *Le groupe de Brauer-Grothendieck en géométrie diophantienne*, Actes du Congrès Intern. Math. (Nice, 1970), Gauthier-Villars, Paris, 1971, Tome 1, pp. 401–411.
36. —, *Formes cubiques: algèbre, géométrie, arithmétique*, Izdat. "Nauka", Moscou, 1972; trad. ang. North-Holland, Amsterdam-London, 1974, 2^{ème} éd., 1986.
37. Yu. I. Manin and M. A. Tsfasman, *Variétés rationnelles: algèbre, géométrie, arithmétique*, Uspekhi Mat. Nauk **41** (1986), 43–94=Russian Math. Surveys **41** (1986), 51–116.
38. A. S. Merkur'ev and A. A. Suslin, *K-cohomologie des variétés de Severi-Brauer et l'application de norme résiduelle*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), 1011–1046=Math. USSR-Izv. **21** (1983), 307–340.
39. P. Salberger, *K-theory of orders and their Brauer-Severi schemes*, Thesis, Dept. of Math., Univ. of Göteborg, 1985.
40. —, *Galois descent and class groups of orders*, Orders and Their Applications (I. Reiner and K. W. Roggenkamp, eds.), Lecture Notes in Math., vol. 1142, Springer-Verlag, 1985.
41. —, *Sur l'arithmétique de certaines surfaces de Del Pezzo*, C. R. Acad. Sci. Paris **303** (1986), 1273–1276.
42. D. J. Saltman, *Generic Galois extensions and problems in field theory*, Adv. in Math. **43** (1982), 250–283.
43. —, *Retract rational fields and cyclic Galois extensions*, Israel J. Math. **47** (1984), 165–215.

44. —, *Noether's problem over an algebraically closed field*, Invent. Math. **77** (1984), 71–84.
45. —, *The Brauer group and the centre of generic matrices*, J. Algebra **97** (1985), 53–67.
46. —, *Multiplicative field invariants*, J. Algebra **106** (1987), 221–238.
47. J.-J. Sansuc, *Groupe de Brauer et arithmétique des groupes algébriques linéaires sur un corps de nombres*, J. Reine Angew. Math. **327** (1981), 12–80.
48. —, *A propos d'une conjecture arithmétique sur le groupe de Chow d'une surface rationnelle*, Sémin. de Théorie des Nombres de Bordeaux, 1981–1982, Exp. 33.
49. B. Segre, *The rational solutions of homogeneous cubic equations in four variables*, Math. Notae **11** (1951), anno II, 1–68.
50. A. N. Skorobogatov, *Invariants birationnels des surfaces rationnelles*, Mat. Zametki **39** (1986), 736–746=Math. Notes **39** (1986), 404–408.
51. R. G. Swan, *Invariant rational functions and a problem of Steenrod*, Invent. Math. **7** (1969), 148–158.
52. —, *Noether's problem in Galois theory* (Bhama Srinivasan and Judith Sally, eds.), Springer-Verlag, 1983, pp. 21–40.
53. V. E. Voskresenskii, *Propriétés birationnelles des groupes algébriques linéaires*, Izv. Akad. Nauk SSSR Ser. Mat. **34** (1970), 3–19=Math. USSR-Izv. **4** (1970), 1–17.
54. —, *Sur la question de la structure du sous-corps des invariants d'un groupe cyclique d'automorphismes du corps $\mathbb{Q}(x_1, \dots, x_n)$* , Izv. Akad. Nauk SSSR **34** (1970), 366–375=Math. USSR-Izv. **4** (1970), 371–380.
55. —, *Quelques problèmes de géométrie birationnelle des tores algébriques*, Proc. Internat. Congr. Math. (Vancouver, 1974), vol. 1, Canad. Math. Congress, Montreal, Quebec, 1975, pp. 343–347.
56. —, *Tores algébriques*, "Nauka", Moscou, 1977.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE, MATHÉMATIQUES, BÂTIMENT 425, UNIVERSITÉ DE PARIS-SUD, F-91405, ORSAY, FRANCE

Discrete Reflection Groups in Lobachevsky Spaces and Algebraic Surfaces

V. V. NIKULIN

1. Introduction. The report of E. Vinberg at the 1983 International Congress of Mathematicians [61] was devoted to discrete reflection groups in Lobachevsky spaces. In this report we will review some new results in this area and also some of their applications to algebraic geometry which were not considered in Vinberg's report.

The reflection groups (discrete) in Riemannian manifolds of constant positive curvature (finite reflection groups) and of zero curvature (Euclidean reflection groups) were classified by Coxeter [8] (see also [4]). Although their theory has been completed, they continue to attract attention for their numerous applications to different areas of mathematics, for instance to the theories of Lie algebras and groups, algebraic groups, and singularities.

A similar situation occurs for the reflection groups in Lobachevsky spaces (= hyperbolic reflection groups = HRG). However, in this case the classification of such groups is far from being complete and represents a very difficult and interesting problem. This is the reason why we shall discuss this problem first. The importance of HRG has been revealed recently in many theories. Here we restrict ourselves only to its application to the study of certain classes of algebraic surfaces.

2. Classification of crystallographic hyperbolic reflection groups (CHRG). The description of HRG can be reduced to the description of their fundamental polyhedra \mathcal{M} , which are convex polyhedra in a Lobachevsky space \mathcal{L} , the facial angles of which (the angles between the faces of highest dimension) are of the form $\pi/n, n \in \mathbb{N}$. A reflection group W is generated by all reflections into faces of \mathcal{M} . As usual, we assume that \mathcal{M} is of finite volume; then W is said to be a *crystallographic group* (CHRG). In this case \mathcal{M} is a convex closed polyhedron with finitely many faces.

At the present time, the fundamental polyhedra \mathcal{M} of CHRG have been completely classified only if $\dim \mathcal{L} = 2$ (Poincaré [48] and Dyck [14]) or $\dim \mathcal{L} = 3$ (Andreev [1, 2]).

However, recently some classification results were obtained in arbitrary dimension. The fundamental approach to this problem consists of studying some general properties of CHRG fundamental polyhedra \mathcal{M} . At present, three such properties have been found, which we now state. They are satisfied for arbitrary convex closed polyhedra.

THEOREM A (*on the narrow face of a polyhedron*). *For any convex closed polyhedron Q in Lobachevsky space (of curvature -1) there is a (narrow) face \mathcal{N}_{\min} such that the distance $\rho(H(\mathcal{A}), H(\mathcal{B})) < \operatorname{arccosh} 7$ for any faces \mathcal{A} and \mathcal{B} satisfying $H(\mathcal{A}) \cap H(\mathcal{N}_{\min}) \neq \emptyset$, $H(\mathcal{B}) \cap H(\mathcal{N}_{\min}) \neq \emptyset$, where $H(\mathcal{X})$ is the hyperplane containing the face \mathcal{X} .*

THEOREM B (*the delicate theorem on the narrow face of a polyhedron*). *For any convex closed polyhedron Q in the n -dimensional Lobachevsky space (of curvature -1) there are faces $\mathcal{A}_1, \dots, \mathcal{A}_{n+1}$ such that (a) $\rho(H(\mathcal{A}_i) \cap H(\mathcal{A}_j)) < \operatorname{arccosh} 31$, $1 \leq i, j \leq n+1$; (b) $H(\mathcal{A}_1), \dots, H(\mathcal{A}_{n+1})$ are in general position, i.e., they do not have common points and no hyperplane is orthogonal to all of them; (c) the set $H(\mathcal{A}_1), \dots, H(\mathcal{A}_{n+1})$ cannot be decomposed into the union of two subsets such that any hyperplane from one of them is orthogonal to any hyperplane from the other one.*

A polyhedron Q is said to be *simple* (resp. *almost simple*) if it is a cone over a simplex (resp. a simple polyhedron) in a neighborhood of any vertex. It is well known that a bounded fundamental polyhedron of CHRG is simple, and an unbounded one (with some vertices at infinity) is almost simple.

THEOREM C (*a combinatorial bound*). *Let Q be an almost simple n -dimensional convex polyhedron. Then the average number $\alpha_k^l(Q)$ of l -dimensional faces of k -dimensional faces of Q has the following bound:*

$$\alpha_k^l(Q) < \binom{n-l}{n-k} \left(\binom{\lfloor n/2 \rfloor}{l} + \binom{n - \lfloor n/2 \rfloor}{l} \right) / \left(\binom{\lfloor n/2 \rfloor}{k} + \binom{n - \lfloor n/2 \rfloor}{k} \right)$$

if $l \leq k \leq (n+1)/2$. Here the right side is decreasing and tends to $\binom{k}{l} 2^{k-l}$ (it is the number of l -dimensional faces of the k -dimensional cube) when n tends to infinity.

Theorems A and B were proven in [40, 41]. The proof of Theorem A is simple, and we believe that the result is known to specialists. The proof of Theorem B is much more complicated; it is hard to find the faces satisfying the additional delicate condition (c). In the case of simple polyhedra, Theorem C was proven in [41]. In the case of almost simple polyhedra, the theorem was proven recently by Khovanskii [25, 26, 27]. In both cases the properties of some combinatorial polynomials attached to simple polyhedra were used (see [52] and also [9]).

Now let us turn our attention to applications of Theorems A, B, and C to CHRG. First, let us recall that there are arithmetic and nonarithmetic CHRG. Examples of nonarithmetic CHRG are known in dimensions 2, 3, 4, and 5 (see

[32, 33, 55]). The class of arithmetic CHRG is especially important for applications.

We have the following description of arithmetic CHRG (see [55]). Let \mathbf{K} be a totally real algebraic number field (the ground field), \mathbf{O} be the ring of integers in \mathbf{K} , and $N = [\mathbf{K} : \mathbf{Q}]$. Let S be a lattice over \mathbf{O} , i.e., a projective \mathbf{O} -module of finite rank equipped with a nondegenerate symmetric bilinear form with values in \mathbf{O} . A lattice S is called *hyperbolic* if for some imbedding $\sigma^{(+)}: \mathbf{K} \rightarrow \mathbf{R}$ the form $\Phi = S \otimes_{\mathbf{O}} \mathbf{R}$ over \mathbf{R} is hyperbolic (i.e., of signature $(1, n)$) and for other imbeddings $\sigma: \mathbf{K} \rightarrow \mathbf{R}$, $\sigma \neq \sigma^{(+)}$, the form $S \otimes_{\mathbf{Q}} \mathbf{R}$ is definite. For a hyperbolic lattice S of $\text{rank}(S) \geq 3$, we denote by $\mathcal{L}(S)$ the Lobachevsky space of dimension $\text{rank}(S) - 1$ associated to the form Φ . Let $O(S)$ be the automorphism group of S and $O(S) = O'(S) \times \{\pm 1\}$. The group $O'(S)$ is a discrete subgroup of the group of motions of $\mathcal{L}(S)$ and has fundamental domain in $\mathcal{L}(S)$ of finite volume. Let $W(S)$ be the subgroup of $O'(S)$ generated by all reflections (as motions of $\mathcal{L}(S)$) in $O'(S)$. If $W(S)$ is of finite index in $O'(S)$, $W(S)$ is an arithmetic group of motions of $\mathcal{L}(S)$. Every arithmetic CHRG is a subgroup of finite index in $W(S)$. Therefore, the classification of maximal arithmetic CHRG is reduced to the classification of hyperbolic lattices S with $[O'(S):W(S)] < \infty$. We call them *reflective*. The reflectivity is obviously preserving if one multiplies the form of the lattice by any $k \in \mathbf{K}$ and takes the similar lattice.

Theorems A, B, and C were applied to arithmetic CHRG in [40, 41] for the proof of the following result:

THEOREM 1. (a) *If $n = \dim \mathcal{L} \geq 2$ and $N = [\mathbf{K} : \mathbf{Q}]$ are fixed, the number of ground fields \mathbf{K} and the number of similarity classes of reflective lattices (and, consequently, maximal arithmetic CHRG) are finite. (This follows directly from Theorem B).*

(b) *If N is sufficiently large, then $n \leq 9$ (this follows directly from Theorems A and C for simple polyhedra and some arithmetic results [41]).*

Later, applying Theorem C in the case of simple polyhedra and $k = 2$, Vinberg proved the following result [59, 62, 63]:

THEOREM 2. *For CHRG with bounded fundamental polyhedron*

$$n = \dim \mathcal{L} < 30.$$

Recently, using Khovanskiĭ's proof of Theorem C for almost simple polyhedra and also the ideas of Vinberg [59, 62, 63] and the author [41], Khovanskiĭ [27] and Prokhorov [49] have proved the following result:

THEOREM 3. *For CHRG with unbounded fundamental polyhedron*

$$n = \dim \mathcal{L} \leq 995.$$

We have to point out that there is a purely arithmetic proof that $\dim \mathcal{L} < 30$ for arithmetic CHRG with unbounded fundamental polyhedron (see [62] and also [39]).

A bound for $\dim \mathcal{L}$, which follows from Theorems 1(b), 2, and 3, is undoubtedly a central result of the theory of CHRG. Unfortunately, Theorem C, which is used for the proof of these theorems, gives information only about average values of some parameters of fundamental polyhedra of CHRG, and it does not help in classifying CHRG in small dimension. One of the most important problems of the theory of CHRG is finding a method of bounding CHRG dimension which does not have this deficiency.

Recently, we have found a method for bounding CHRG dimension which does not use Theorem C but is based on Theorem A. Let us describe this.

Let \mathcal{M} be a CHRG fundamental polyhedron and V be a vertex of its narrow face $\mathcal{N} = \mathcal{N}_{\min}$ (in the notation of Theorem A). Let r_i , $i \in I$, be the edges of \mathcal{N} which contain V , $V_i \neq V$ be their other ends, and \mathcal{N}_i be the faces of \mathcal{M} such that $r_i \cap \mathcal{N}_i = V_i$.

THEOREM 4. (a) *If V is an infinite vertex, there exist two faces \mathcal{N}_i and \mathcal{N}'_i , $i, i' \in I$, such that $\rho(H(\mathcal{N}_i), H(\mathcal{N}'_i)) > f(\dim \mathcal{L})$.*

(b) *If V is a finite vertex, the previous bound can be replaced by*

$$\rho(H(\mathcal{N}_i), H(\mathcal{N}'_i)) > g(l),$$

where l is the maximum of the lengths of connected components of the Coxeter diagram of V (of faces of \mathcal{M} containing V).

(c) *After repeating the definition of \mathcal{N}_i for a finite V k times (i.e., defining the edges r_{ij} coming from V_i along \mathcal{N} , and so on), one obtain the faces $\mathcal{N}_{i_1 \dots i_k}$, $\mathcal{N}'_{i'_1 \dots i'_k}$ satisfying an estimate $\rho(H(\mathcal{N}_{i_1 \dots i_k}), H(\mathcal{N}'_{i'_1 \dots i'_k})) > h(t)$, where t is the number of connected components of the Coxeter diagram of V .*

Here f, g, h are some functions the value of which tend to infinity (when the argument does) and can be explicitly read off from the Coxeter diagram of V .

Since $l \leq \dim \mathcal{L}$, $t \leq \dim \mathcal{L}$, for large $\dim \mathcal{L}$, this theorem contradicts the definition of \mathcal{N}_{\min} and gives a bound for $\dim \mathcal{L}$.

Some explicit formulae play important roles in the proof of Theorem 4. Let us give an example of one of them.

We assume that \mathcal{L} is defined by the pseudo-Euclidean space $E^{1,n}$ of signature $(1, n)$ and by the halfcone V^+ of positive elements $x \in E^{1,n}$, $x^2 = x \cdot x > 0$ (see [55]). Then \mathcal{L} is the set of all rays $[x] \in V^+/\mathbf{R}^+$, $x \in V^+$, $\dim \mathcal{L} = n$. Every halfspace H^+ of \mathcal{L} is equal to some $H_\delta^+ = \{[x] \in \mathcal{L} | x \cdot \delta \geq 0\}$, where $\delta \in E^{1,n}$ with $\delta^2 = -2$. The element δ is orthogonal to the face H_δ of the halfspace H_δ^+ .

Let V be an infinite vertex of type \tilde{A}_{n-1} (i.e., its Coxeter diagram is of type \tilde{A}_{n-1}) belonging to the face $\mathcal{N} = \mathcal{N}_{\min}$ of a fundamental polyhedron \mathcal{M} . Let $(\delta_1, \dots, \delta_n) \subset E^{1,n}$ be the orthogonal vectors to the faces of \mathcal{M} which contain V , $\delta_i^2 = -2$. They are described by the diagram \tilde{A}_{n-1} (Figure 1), i.e., $\delta_i \cdot \delta_j = 1$ if i, j are connected by an edge of the diagram, and $\delta_i \cdot \delta_j = 0$ otherwise. The vertex V is defined by the ray $[c]$, $c = \delta_1 + \dots + \delta_n$. Any element $x \in E^{1,n}$ with $x \cdot c \neq 0$ is uniquely determined by the row $x = (k_1, \dots, k_n; x^2)$, where

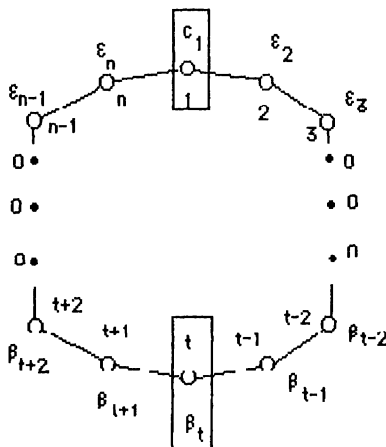


Figure 1

$k_i = x \cdot \delta_i$. Let $y \in E^{1,n}$ with $y \cdot c \neq 0$ be similarly determined by the row $y = (m_1, \dots, m_n; y^2)$. We have the formula

$$\begin{aligned} x \cdot y &= \frac{1}{2} x^2 \left(\sum m_i / \sum k_i \right) + \frac{1}{2} y^2 \left(\sum k_i / \sum m_i \right) \\ &\quad - \frac{1}{2} \varphi(k, k) \left(\sum m_i / \sum k_i \right) \\ &\quad - \frac{1}{2} \varphi(m, m) \left(\sum k_i / \sum m_i \right) + \varphi(k, m), \end{aligned} \quad (*)$$

where

$$\varphi(k, m) = \left(\frac{1}{2n} \right) \sum_{1 \leq i < j \leq n} (k_i m_j + k_j m_i) (j - i) (n + 1 - j + i),$$

$k = (k_1, \dots, k_n)$, $m = (m_1, \dots, m_n)$.

Let us consider the edge r_1 coming out of V and containing all faces of \mathcal{M} which are orthogonal to some elements from the set $\{\delta_2, \dots, \delta_n\}$. Let \mathcal{N}_1 be the face of \mathcal{M} containing the second end V_1 of the edge r_1 . For $n \geq 18$ the face \mathcal{N}_1 is orthogonal to an element $e_1 \in E^{1,n}$ defined by the row

$$e_1 = (\varepsilon_1, \varepsilon_2, \varepsilon_3, 0, \dots, 0, \varepsilon_{n-1}, \varepsilon_n; -2),$$

where $0 \leq \{\varepsilon_2, \varepsilon_3, \varepsilon_{n-1}, \varepsilon_n\} \leq \sqrt{2}$, $\varepsilon_1 \geq 0$, and $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_{n-1} + \varepsilon_n \geq 1$. Here we are using the fact that $(e_1, \delta_2, \delta_3, \dots, \delta_{n-1})$ have an elliptic or parabolic Coxeter diagram and $(e_1, \delta_1, \delta_2, \delta_3, \dots, \delta_n)$ generate the hyperbolic space $E^{1,n}$. Similarly assume \mathcal{N}_t is defined by the edge r_t which contains all faces of \mathcal{M} orthogonal to some elements of the set $\{\delta_1, \delta_2, \dots, \hat{\delta}_t, \dots, \delta_n\}$. Then it is orthogonal to the element

$$e_t = (0, \dots, 0, \beta_{t-2}, \beta_{t-1}, \beta_t, \beta_{t+1}, \beta_{t+2}, 0, \dots, 0; -2),$$

where $0 \leq \{\beta_{t-2}, \beta_{t-1}, \beta_{t+1}, \beta_{t+2}\} \leq \sqrt{2}$, $\beta_t \geq 0$, and $\beta_{t-2} + \beta_{t-1} + \beta_t + \beta_{t+1} + \beta_{t+2} \geq 1$.

We may assume that the face $\mathcal{N} = \mathcal{N}_{\min}$ is orthogonal to δ_i , $i \neq 1, t$. Since $\mathcal{N}_1 \cap \mathcal{N} \neq \emptyset$, $\mathcal{N}_t \cap \mathcal{N} \neq \emptyset$, $\mathcal{N} \cap H_{\delta_1} \neq \emptyset$, $\mathcal{N} \cap H_{\delta_t} \neq \emptyset$, we obtain that $2 \cosh(\rho(H_{\delta_1}, H_{e_1})) = \delta_1 \cdot e_1 = \varepsilon_1 < 14$, and similarly, $\beta_t < 14$.

For $t = \lfloor n/2 \rfloor$ we obtain from formula (*) that $2 \cosh(\rho(H(\mathcal{N}_1), H(\mathcal{N}_t))) = (e_1, e_t) \geq u + n/8$, where u is an absolute constant. This proves Theorem 4 in the case where the vertex V is of type \tilde{A}_{n-1} . On the other hand, since $\mathcal{N} = \mathcal{N}_{\min}$, $(e_1, e_t) < 14$, this gives a bound for $n = \dim \mathcal{L}$. More precise calculation shows that $n < 140$ in this case.

For small n , when Theorem 4 does not lead to contradiction, our method gives us rather much information about fundamental polyhedra \mathcal{M} of CHRG in a neighborhood of a vertex V of \mathcal{N}_{\min} . In the arithmetic case over a fixed ground field, the situation is discrete and there is a hope to find all possible pieces of the fundamental polyhedron \mathcal{M} in a neighborhood of \mathcal{N}_{\min} . Moreover, Theorem 1(a) and the bound for the dimension n follows from similar facts (Theorems A and B).

We see that each of Theorems A and C gives a bound for the dimension. However, sometimes it is useful to apply them both at the same time (as in the proof of Theorem 1(b)). For example, if the face \mathcal{N}_{\min} of \mathcal{M} does not have infinite vertices, we can apply Theorem C to \mathcal{N}_{\min} in the case of simple polyhedra and Vinberg's arguments from [59, 62, 63]. Or, we can apply Theorem 4(a) to \mathcal{N}_{\min} . In this way we can obtain that $\dim \mathcal{L} < 300$. The last inequality is better than the one obtained by Khovanskii and Prokhorov.

The next problem is fundamental for the classification of arithmetic CHRG.

PROBLEM. *Are the degrees of the ground fields of arithmetic CHRG bounded if $\dim \mathcal{L} < 9$?*

3. A generalization of crystallographic hyperbolic reflection groups.

We believe that the above results on CHRG can be generalized to the following two classes of groups.

Let G be a discrete group of motions of a Lobachevsky space \mathcal{L} with fundamental polyhedron of finite volume. We say that G is a generalized crystallographic hyperbolic reflection group (GenCHRG) if there is a normal subgroup $W \triangleleft G$ generated by reflections such that the factor group $A = G/W$, considered as a group of symmetries of a certain fundamental polyhedron \mathcal{M} of W , contains a subgroup A' of finite index which leaves invariant a proper subspace \mathcal{L}' of \mathcal{L} . For example, if \mathcal{L}' is spanned by a finite point of \mathcal{L} , then A is a finite group and W is a CHRG. If \mathcal{L} is spanned by an infinite point, then A has a finitely generated abelian subgroup of finite index. The group A and its action on \mathcal{L} are especially interesting in applications (in particular to $K3$ -surfaces). In the same way we can define arithmetic GenCHRG and generalized reflective lattices. In [39] there are many examples of arithmetic GenCHRG which appear in connection with $K3$ -surfaces. J. Conway pointed out in [6] that the automorphism group $O'(S)$ of the hyperbolic lattice $S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \perp K$, where K is the Leech lattice,

is a GenCHRG. In this case \mathcal{L}' is an infinite point and A contains an abelian subgroup $A' \cong \mathbb{Z}^{24}$ of finite index.

We think that all results from §2 can be generalized to GenCHRG. In [39] one can find some qualitative results on GenCHRG in the case of hyperbolic lattices S over \mathbb{Z} , $G = O'(S)$ and $W = W^{(2)}(S)$. The latter denotes the subgroup of $O'(S)$ generated by reflections with respect to all elements in S of square -2 .

PROBLEM. *Is there a generalization of the results of §2 to discrete reflection groups in the complex ball?*

4. Hyperbolic reflection groups and automorphisms of algebraic surfaces. Here we consider some applications of HRG to automorphisms of algebraic surfaces over \mathbb{C} of three types: (i) $K3$ -surfaces, (ii) DPN-surfaces (one of the possible generalizations of Del Pezzo surfaces), (iii) Enriques surfaces (this is a special case of DPN-surfaces). As a result of the classification of algebraic surfaces in classical works of the Italian school of algebraic geometry at the beginning of this century, the problem of describing the group of birational automorphisms of an algebraic surface (up to finite groups) remained open only for $K3$ and Enriques surfaces. Except for Enriques surfaces, DPN surfaces are rational. A description of their automorphism groups is interesting since we know very little about biregular automorphisms of rational surfaces.

In all three cases (i)–(iii), the same objects appear. They are a hyperbolic lattice S over \mathbb{Z} isomorphic to a sublattice of the Picard lattice of the surface, a reflection group $W \subset O'(S)$ in the Lobachevsky space $\mathcal{L}(S)$ with a fixed fundamental polyhedron $\mathcal{M}_{\text{geom}}$ containing the polarization ray of the surface, the subset $P = P(\mathcal{M}_{\text{geom}})$ of all primitive elements $\delta \in S$ orthogonal to the face hyperplanes of $\mathcal{M}_{\text{geom}}$ which point outward (they have negative square). Let $O'(S)_W$ be the normalizer of W in $O'(S)$, $A \subset O'(S)_W$ the automorphism group of $\mathcal{M}_{\text{geom}}$. The group $O'(S)_W$ is the semidirect product $W \ltimes A$. In each case (i)–(iii) we have the following theorem, the first statement of which is a consequence of the Riemann-Roch theorem for surfaces (and is valid in arbitrary characteristic), while the second one is a consequence of the deep Global Torelli theorem for $K3$ -surfaces, due to Piatetski-Shapiro and Shafarevich [47] (see also [12, 36, 44] in cases (ii) and (iii)).

THEOREM 5. (a) *The subgroup $O'(S)_W \subset O'(S)$ is of finite index. P is the set of all divisor classes of “exceptional curves” of the surface. The action of the automorphism group G of the surface on the Picard lattice preserves the sublattice S ; the kernel of its representation in $O(S)$ is finite and its image lies in A .*

(b) *The image of G in A is of finite index in $A \simeq O'(S)_W/W$. In particular, if $\text{rank}(S) \geq 3$, G is finite if and only if W is CHRG in $\mathcal{L}(S)$. In this case the set P of “exceptional divisor classes” is finite.*

Since $\mathcal{M}_{\text{geom}}$ is uniquely defined up to W -action, W describes the intersection geometry of exceptional curves and the action of the automorphism group G of the surface on the lattice S and the geometry (up to finite groups).

Now let us describe precisely our cases (i)–(iii), the corresponding automorphism groups G , the sets of exceptional curves, the lattice S , the group W , and some known results about their computations.

(i) *K3-surfaces*. Let X be a K3-surface. The group $G = \text{Aut}(X)$. An “exceptional curve” is any nonsingular rational curve on X (its self-intersection is equal to -2). The lattice $S = \text{Pic}(X)$. The reflection group $W = W^{(2)}(S)$ is the group of 2-reflections of S ; it is generated by all 2-reflections $s_\delta: x \mapsto x + (x \cdot \delta)\delta$, where $\delta, x \in S$ and $\delta^2 = -2$. In this case Theorem 5 was proven in [47]. Theorems 1(a) and 3 show that there are only a finite number of hyperbolic lattices over \mathbf{Z} with $\text{rank } S \geq 3$ and finite group $\text{Aut}(X) \approx A \cong O'(S)/W^{(2)}(S)$ (\approx denotes an isomorphism up to finite groups). These lattices are called 2-reflective. All such lattices have been found. If $\text{rank } S \geq 5$ and $\text{rank } S = 3$, it has been done in papers [38, 39, 43] (see also [11]); if $\text{rank}(S) = 4$ it has been done by E. Vinberg (unpublished). In the notation of [38, 39] we have the following 17 lattices of rank $S = 4$:

$$\langle 8 \rangle \oplus (A_1)^3; \quad U(k) \oplus (A_1)^2, \quad k = 1, 2, 3, 4; \quad U(k) \oplus A_2, \quad k = 1, 2, 3, 6;$$

$$\begin{bmatrix} 0 & -3 \\ -3 & 2 \end{bmatrix} \oplus A_2; \quad \langle 4 \rangle \oplus \langle -4 \rangle \oplus A_2; \quad \langle 4 \rangle \oplus A_3;$$

$$\begin{bmatrix} 2 & -1 & -1 & -1 \\ -1 & -2 & 0 & 0 \\ -1 & 0 & -2 & 0 \\ -1 & 0 & 0 & -2 \end{bmatrix}; \quad \begin{bmatrix} 12 & -2 & 0 & 0 \\ -2 & -2 & -1 & 0 \\ 0 & -1 & -2 & -1 \\ 0 & 0 & -1 & -2 \end{bmatrix}.$$

For every 2-reflective lattice S , one can describe the fundamental polyhedron \mathcal{M} of $W^{(2)}(S)$ (see [39, 43] for the case $\text{rank } S \neq 4$). The method of finding reflective lattices of rank ≥ 6 is the following. We find those K3-surfaces with $\text{rank}(S) \geq 6$ which have elliptic pencils with only finitely many translations. This implies that the automorphism group of such a K3-surface is finite. In the lattice language this means that for every primitive isotropic vector $c \in S$ the negative definite lattice $(\mathbf{Z}c)_S^\perp / \mathbf{Z}c$ is generated, up to finite index, by elements with square -2 . The application of this method requires rather delicate lattice-theoretical arguments and is based on the technique of discriminant forms which was developed in [37].

(ii) *DPN-surfaces*. Here we consider pairs (X, σ) , where X is a K3-surface and σ is an involution of X for which $\sigma^*(\omega) = -\omega$, where ω is a regular 2-form on X .

In this case the fixed-point set X^σ of σ is empty or a nonsingular curve on X , the factor surface $Y = X/\{1, \sigma\}$ is a nonsingular surface, the projection

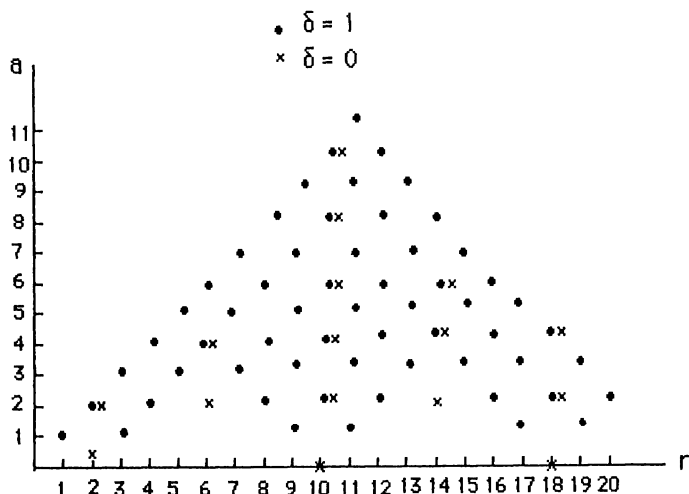


Figure 2

$\pi: X \rightarrow Y$ is a double covering of Y branched along a nonsingular curve $C = \pi(X^\sigma) \in |-2K_Y|$, where K_Y is the canonical class of Y . The surface Y is an Enriques surface if $X^\sigma = \emptyset$, and is a rational surface otherwise.

Let us consider all pairs (Y', C') obtained from all pairs (Y, C) , as above, by contraction of exceptional curves of the first kind. It is very easy to prove that Y' is a nonsingular surface and $C' \in |-2K_{Y'}|$ is a curve with only simple singularities (i.e., ADE-singularities). Such pairs are called DPN-pairs; the corresponding surfaces Y are called DPN-surfaces. We may consider DPN-surfaces as one of many possible generalizations of Del Pezzo surfaces (see [10, 13, 34]). It is very easy to show that a minimal model of a DPN-surface is \mathbf{P}^2 or \mathbf{F}_n , $n = 0, 2, 3, 4$. The classification of DPN-surfaces and DPN-pairs, the description of their exceptional curves (i.e., irreducible rational curves with negative squares), and their automorphism groups are reduced to the description of the pairs (X, σ) from above.

Let $S_X = \text{Pic}(X)$ be the Picard lattice of X and $(S_X)_\pm = \{x \in S_X: \sigma^*(x) = \pm x\}$. In our case $S = (S_X)_+$. The isomorphism class of the lattice S is the fundamental invariant of the DPN-pair (X, σ) . It is determined by a triplet (r, a, δ) , where $r = \text{rank } S$, $S^*/S \cong (\mathbf{Z}/2\mathbf{Z})^a$, $\delta = 0$ if $x^{*2} \in \mathbf{Z}$ for any $x^* \in S^* = \text{Hom}(S, \mathbf{Z})$, and 1 otherwise. All possible triplets (r, a, δ) are found and are shown in Figure 2.

For any given triplet (r, a, δ) , the corresponding pairs (X, σ) and their factors (Y, C) give us connected smooth families of nonpolarized surfaces

$$\begin{array}{ccc} \sigma \hookrightarrow \mathcal{X}_{(r,a,\delta)} & \rightarrow & \mathcal{G}_{(r,a,\delta)} \supset \mathcal{C}_{(r,a,\delta)} \\ & \searrow & \nearrow \\ & \mathcal{M}_{(r,a,\delta)} & \end{array}$$

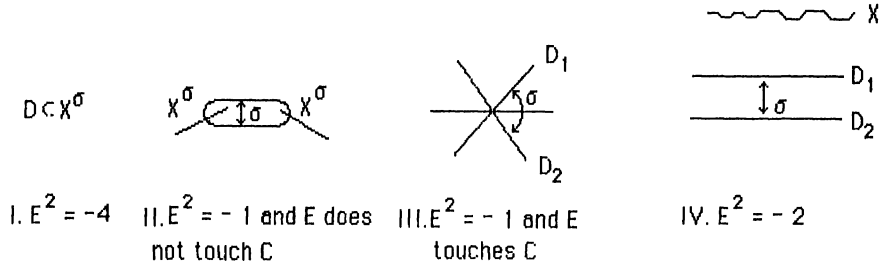


Figure 3

It can be shown that $\dim \mathcal{M}_{(r,a,\delta)} = 20 - r$. The curve $X^\sigma \cong C$ has the following components: If $(r, a, \delta) \neq (10, 10, 0)$ and $(10, 8, 0)$, then $X^\sigma = C_g + E_1 + \dots + E_k$, where C_g is a curve of genus $g = \frac{1}{2}(22 - r - a)$, E_i are rational curves, $k = \frac{1}{2}(r - a)$; if $(r, a, \delta) = (10, 10, 0)$, $X^\sigma = \emptyset$; if $(r, a, \delta) = (10, 8, 0)$, $X^\sigma = C_1 + C'_1$, where C_1, C'_1 are elliptic curves. Moreover, X^σ is divisible by 2 in $H^2(X; \mathbf{Z})$ if and only if $\delta = 0$. The divisor classes of the components of X^σ can be calculated via a fundamental polyhedron \mathcal{M} of $W^{(2)}(S)$ containing a polarization ray of X . Now let us consider "general" pairs (X, σ) of the family $\mathcal{M}_{(r,a,\delta)}$, i.e., the pairs with $S_X = (S_X)_+ = S$. In this case an "exceptional curve" means an exceptional curve of the $K3$ -surface X (see case (i)). Their images with respect to the morphism π give all exceptional curves on Y . Each of them is an exceptional curve of the first kind except those which are equal to $E = \pi(D)$ with $E^2 = -4$ and D any rational component of X^σ . In Theorem 5, $G = \text{Aut}(X, \sigma) = \text{Aut}(X)$; hence, $\text{Aut}(X)/\{1, \sigma\} \cong \text{Aut}(Y, C)$ and $W = W^{(2)}(S)$. These results are of arithmetic nature and were obtained in [39]. They give the classification results which are presently known for DPN-surfaces.

Now let us consider an arbitrary pair (X, σ) of the family $\mathcal{M}_{(r,a,\delta)}$. By definition, an exceptional curve of the pair (X, σ) is a curve $D = \pi^{-1}(E)$, where E is an exceptional curve on Y . The curve D may be of one of the 4 types presented in Figure 3. For types I and II, D is an exceptional curve on X ; for types III and IV, $D = D_1 + D_2$, where D_1, D_2 are exceptional curves on X .

The group G is equal to $\text{Aut}(X, \sigma)$, therefore $G/\{1, \sigma\} = \text{Aut}(Y, C)$. To describe the group W we need some notations. Let

$$\Delta_{\pm}^{(4)} = \{\delta_{\pm} \in (S_X)_{\pm} : \delta_{\mp}^2 = -4 \text{ and } \exists \delta_{\mp} \in (S_X)_{\mp} \text{ with } \delta_{\mp}^2 = -4, (\delta_{\pm} + \delta_{\mp})/2 \in S_X\},$$

$$\Delta_+^{(2)} = \{\delta_+ \in (S_X)_+ : \delta_+^2 = -2\},$$

$$\Delta_{+t}^{(2)} = \{\delta_+ \in \Delta_+^{(2)} : \exists \delta_- \in (S_X)_- \text{ with } \delta_-^2 = -6, (\delta_+ + \delta_-)/2 \in S_X\},$$

$$\Delta_-^{(6)} = \{\delta_- \in (S_X)_- : \delta_-^2 = -6 \text{ and } \exists \delta_+ \in \Delta_{+t}^{(2)} \text{ with } (\delta_+ + \delta_-)/2 \in S_X\}.$$

The reflection group $W \subset O'(S)$ is generated by all reflections with respect to elements from $\Delta_+^{(2)} \cup \Delta_+^{(4)}$. The set $P = P(\mathcal{M}_{\text{geom}})$ of the divisor classes of exceptional curves of (X, σ) is divided into four subsets $P_I, P_{II}, P_{III}, P_{IV}$

corresponding to the four types from Figure 3. We have

$$P_{\text{III}} = P \cap \Delta_{+t}^{(2)}, \quad P_{\text{II}} = (P \cap \Delta_+^{(2)}) - (P_1 \cup P_{\text{III}}), \quad P_{\text{IV}} = P \cap \Delta_+^{(4)}.$$

Let $W^{(2,4)}(S) \subset O'(S)$ be the reflection group generated by all reflections with respect to elements $\delta \in S$ with $\delta^2 = -2$ or -4 . Let $M_0 \subset M_{\text{geom}}$ be a fundamental polyhedron of this group and let $M_{\text{geom}} = \mathcal{W}M_0$, where $\mathcal{W} \subset W^{(2,4)}(S)$ is the set of representatives of the set of cosets $W \backslash W^{(2,4)}(S)$. The set \mathcal{W} is a generalization of the Weyl group introduced for Del Pezzo surfaces to the case of DPN-surfaces (see [34] and compare with calculations in [57, 58]).

To describe all possible groups $\text{Aut}(X, \sigma)$ for a given S (i.e., for a given triplet (r, a, δ)), we introduce some additional invariants of (X, σ) . The generalized root invariant of the pair (X, σ) is the triplet $\tilde{R}(X, \sigma) = (\tilde{K}, \Delta_-^{(4)} \cup \Delta_-^{(6)}, \tilde{\gamma})$. Here $\tilde{K} = [\Delta_-^{(4)} \cup \Delta_-^{(6)}](\frac{1}{2})$ is the negative definite lattice which is spanned by the subset $\Delta_-^{(4)} \cup \Delta_-^{(6)} \subset (S_X)_-$ and the quadratic form of which is multiplied by $\frac{1}{2}$. The subset $\Delta_-^{(4)} \cup \Delta_-^{(6)}$ gives some special set of generators of the lattice \tilde{K} with square -2 and -3 . The last member of the triplet is a homomorphism $\tilde{\gamma}: \tilde{K} \bmod 2 \rightarrow -q_S$ of quadratic forms with values in $\frac{1}{2}\mathbf{Z}/2\mathbf{Z}$. Here $\tilde{K} \bmod 2$ is the abelian group $\tilde{K}/2\tilde{K}$ with the quadratic map given by the formula $\tilde{x}^2 = \frac{1}{2}x^2$. Moreover, q_S denotes the discriminant form [37] of S defined as the abelian group $A_S = S^*/S \cong (\mathbf{Z}/2\mathbf{Z})^a$ with the quadratic map obtained by extending the quadratic form from S to $S^* = \text{Hom}(S, \mathbf{Z}) \supset S$. The map $\tilde{\gamma}$ is defined by the formula

$$\tilde{\gamma}(\delta_- \bmod 2) = \delta_+/2 \bmod S,$$

where $\delta_+ \in S$, $\delta_- \in \tilde{K}$, and $\frac{1}{2}(\delta_+ + \delta_-) \in S_X$. The invariant $\tilde{R}(X, \sigma)$ is defined up to isomorphisms of lattices \tilde{K} and automorphisms of S . It provides us with all the data which we need for Theorem 5. Indeed,

$$\tilde{\gamma}(\Delta_-^{(4)} \bmod 2) = \frac{1}{2}\Delta_+^{(4)} \bmod S, \quad \tilde{\gamma}(\Delta_-^{(6)} \bmod 2) = \frac{1}{2}\Delta_{+t}^{(2)} \bmod S.$$

This shows that there is only a finite number of possible groups $\text{Aut}(X, \sigma)$ (up to finite groups) and of the intersection geometries of exceptional curves of (X, σ) (since $\text{rank}(S) + \text{rank}(\tilde{K}) \leq 20$). Let us mention that $\Delta_+^{(4)}, W$ and hence $\text{Aut}(X, \sigma), P$ are determined by a simpler invariant, that is, the root invariant $R(X, \sigma) = (K, \Delta_-^{(4)}, \gamma)$, where $K = [\Delta_-^{(4)}](\frac{1}{2})$. Its definition is similar to that of $\tilde{R}(X, \sigma)$. The dimension of the moduli space of DPN-pairs with given invariants S and \tilde{R} (resp. R) is equal to $20 - \text{rank } S - \text{rank } \tilde{K}$ (resp. $20 - \text{rank } S - \text{rank } K$).

(iii) *Enriques surfaces.* In this case $(r, a, \delta) = (10, 10, 0)$ or, equivalently, $X^\sigma = \emptyset$ and $Y = X/\{1, \sigma\}$ is an Enriques surface. We also have $\Delta_+^{(2)} = \Delta_-^{(6)} = \emptyset$, $\tilde{R} = R$, $\Delta_-^{(4)}$ is equal to the set $\Delta^{(2)}(K)$ of all elements of K with square -2 . Thus, R is reduced to the pair (K, H) , where $H = \text{Ker } \gamma$. The condition of the finiteness of the automorphism group of all elliptic pencils of Y can be expressed solely in terms of the finite subset $r = \gamma(\Delta^{(2)}(K) \bmod 2) \subset A_S \cong (\mathbf{Z}/2\mathbf{Z})^{10}$ and the discriminant form q_S on A_S [44, Theorem 3]. This

allows us to find all possible root invariants R of Enriques surfaces with finite automorphism groups. Note that this is a problem which is simpler than the corresponding one for $K3$ -surfaces because we are working with the lattices over $\mathbf{Z}/2\mathbf{Z}$ instead of over \mathbf{Z} .

THEOREM 6. *An Enriques surface Y has finite automorphism group if and only if it has one of the following root invariants:*

$$(E_8 \oplus A_1, \{0\}), (D_9, \{0\}), (E_6 \oplus A_4, \{0\}), (D_5 \oplus D_5, \mathbf{Z}/2\mathbf{Z}),$$

$$(A_9 \oplus A_1, \mathbf{Z}/2\mathbf{Z}), (E_7 \oplus A_2 \oplus A_1, \mathbf{Z}/2\mathbf{Z}), (D_8 \oplus A_1 \oplus A_1, (\mathbf{Z}/2\mathbf{Z})^2).$$

Each of these invariants is realized by some Enriques surface.

This result was proven in [44], where the last case was lost in computations. This was pointed out by Kondō [28], who also has proven that the first two invariants give connected families of Enriques surfaces and the remaining ones define a unique Enriques surface. The fifth invariant gives the first example, pointed out by Fano [15] of an Enriques surface with finite automorphism group. The family $(E_8 \oplus A_1, \{0\})$ was found by Dolgachev [12]. In [28] Kondō presents another method of describing Enriques surfaces with finite automorphism group.

All the above results are of arithmetic nature; they use only the interrelationship between the 2-reflection group of a hyperbolic lattice and its involution and do not depend on the geometric nature of these objects. In this way they can be generalized also to $K3$ -surfaces with an involution which can be holomorphic and the identity on the space of holomorphic 2-forms or can be antiholomorphic.

For some other results closely related to this section see [3, 5, 7, 10, 13, 16–20, 31, 35, 50, 51, 60, 64].

5. Hyperbolic reflection groups and real algebraic surfaces. The appearance of reflection groups in real algebraic geometry is a natural phenomenon since the moduli space of real algebraic varieties has a natural boundary, the real discriminant of codimension 1. In our opinion, further progress in the description of the moduli space of real algebraic varieties consists of finding appropriate reflection groups. For the versal deformation of real simple singularities A, D, E , these groups are finite reflection groups (see Looijenga [30]). For real $K3$ -surfaces they are HRG, which is a consequence of the Global Torelli theorem [47] and the surjectivity of the period map for $K3$ -surfaces [29]. (These two results were applied first to the study of topology of real quartic surfaces in \mathbf{P}^3 by Kharlamov [21–23].) Here we will consider the case of $K3$ -surfaces. All our results here are directly related to Hilbert's sixteenth problem (see [23, 65]).

Let us consider a real $K3$ -surface X which is polarized of degree n . The complex conjugation involution σ acts on the lattice $L = H^2(X, \mathbf{Z})$ and the sublattices

$$L_{\pm} = \{x \in L: \sigma^*(x) = \pm x\}$$

are defined. The polarization class $h \in L_-$, and we define the sublattice $L_-^h = [h]_{L_-}^\perp$. There is an orthogonal decomposition

$$L \otimes \mathbf{Q} = (L_+ \otimes \mathbf{Q}) \oplus (L_-^h \otimes \mathbf{Q}) \oplus ([h] \otimes \mathbf{Q}).$$

Let us consider a real holomorphic nonzero 2-form ω ; we have $\sigma^*(\omega) = \bar{\omega}$. We can write $\omega = \omega_+ + \sqrt{-1}\omega_-$, where $\omega_+ \in L_+ \otimes \mathbf{R}$, $\omega_- \in L_-^h \otimes \mathbf{R}$, $\omega_+^2 = \omega_-^2 > 0$. Since the signature of the lattice L is equal to $(3, 19)$, the lattices L_+ and L_-^h are hyperbolic and the period of X is a point $([\omega_+], [\omega_-])$ of the product of two Lobachevsky spaces. Moreover, it follows from the Riemann-Roch theorem that the $[\omega_\pm]$ do not lie in the hyperplanes orthogonal to $\delta_+ \in L_+$ and $\delta_- \in L_-^h$, $\delta_\pm^2 = -2$. After appropriate identification, we may assume that

$$([\omega_+], [\omega_-]) \in (\mathcal{M}_+)^0 \times (\mathcal{M}_-^h)^0,$$

where \mathcal{M}_+ and \mathcal{M}_-^h are fundamental polyhedra of $W^{(2)}(L_+)$ and $W^{(2)}(L_-^h)$, $()^0$ denotes the subset of interior points. Let $A_+ \subset O'(L_+)$ and $A_- \subset O'(L_-^h)$ be the automorphism groups of \mathcal{M}_+ and \mathcal{M}_-^h , and let

$$(A_+ \times A_-)_L = (A_+ \times A_- \times \{\text{id}\}) \cap O(L)$$

be a subgroup of finite index in $A_+ \times A_-$. By the Global Torelli theorem and the surjectivity of the period map, we obtain that the factor space

$$(A_+ \times A_-)_L \backslash (\mathcal{M}_+)^0 \times (\mathcal{M}_-^h)^0$$

is identified (up to codimension 2) with a connected component of the coarse moduli space of real polarized $K3$ -surfaces with fixed action of the complex conjugation involution on the cohomology [37]. Furthermore, let us consider the corresponding component of the Hilbert scheme of real $K3$ -surfaces (embedded by a complete linear system $|kh|$, $k \geq 3$) and complete it by adding all points representing real $K3$ -surfaces with purely imaginary simple singularities. We will denote the obtained component by $\mathcal{H}(\mathbf{R})$. The group $(A_+ \times A_-)_L$ is isomorphic (up to finite groups) to the fundamental group $\pi_1(\mathcal{H}(\mathbf{R}))$ and the period map maps the boundary of $\mathcal{H}(\mathbf{R})$, which is the real discriminant, into the boundary

$$(A_+ \times A_-)_L \backslash (\mathcal{M}_+ \times \mathcal{M}_-^h - (\mathcal{M}_+)^0 \times (\mathcal{M}_-^h)^0).$$

See [39]. This shows that the description of connected components of the moduli space of real polarized $K3$ -surfaces (the coarse projective classification) is reduced to a purely arithmetic problem of classification of triplets (L, σ^*, h) up to isomorphism. Our paper [37] is devoted to this problem. The description of the components with the corresponding finite fundamental group is reduced to the description of 2-reflective lattices which was discussed in §4(i).

Often, it is necessary to consider real $K3$ -surfaces, where in addition to the polarization h some other divisor classes are fixed. More precisely, we are interested in points $([\omega_+], [\omega_-]) \in \mathcal{M}_+ \times \mathcal{M}_-^h$ which correspond to the surfaces with distinguished divisor classes $h = d_0, d_1, \dots, d_k$ generating in the Picard lattice

$S_X = \{s \in L : s \cdot \omega_+ = s \cdot \omega_- = 0\}$ a primitive sublattice S with the given Gram matrix $(a_{ij}) = (d_i \cdot d_j)$ and with an action of σ on S given by

$$\sigma^*(d_0) = -d_0, \quad \sigma^*(d_i) = \sum k_i^j d_j.$$

Thus, we are interested in the points which lie in the product of the Lobachevsky spaces $\mathcal{L}(L_+^{S_+}) \times \mathcal{L}(L_-^{S_-}) \subset \mathcal{L}(L_+) \times \mathcal{L}(L_-^h)$, where $L_+^{S_+} = (S_+)_{L_+}^\perp$, $L_-^{S_-} = (S_-)_{L_-}^\perp$, $S_\pm = S \cap L_\pm$. These subspaces are considered up to natural action of the group $(A_+ \times A_-^h)_L$ on them, and they parametrize the classes of coarse projective classification of real polarized $K3$ -surfaces with the conditions on the Picard lattice given by the matrices (a_{ij}) and (k_i^j) from above. The description of these subspaces is divided into two problems:

(a) A description of isomorphism classes of triplets (L, σ^*, S) , where L is an even unimodular lattice of signature $(3, 19)$, σ^* is an involution of L , $S \subset L$ is a primitive sublattice with $\sigma^*(S) = S$, and the induced action of σ^* on S is defined by the matrices (a_{ij}) and (k_i^j) (this problem is considered in [42] and also 45).

(b) For a given isomorphism class of a triplet (L, σ^*, S) one has to describe the subspaces $\mathcal{L}(L_+^{S_+}) \times \mathcal{L}(L_-^{S_-}) \subset \mathcal{L}(L_+) \times \mathcal{L}(L_-^h)$ intersecting $\mathcal{M}_+ \times \mathcal{M}_-^h$, and obtain one from another by reflections into the faces of \mathcal{M}_+ and \mathcal{M}_-^h which intersect the subspaces $\mathcal{L}(L_+^{S_+})$ and $\mathcal{L}(L_-^{S_-})$ (this is a multidimensional analog of a rational trajectory in a rectangular billiards in a Lobachevsky space). They are considered up to the action of the group $(A_+, A_-)_L$. See [42, 46].

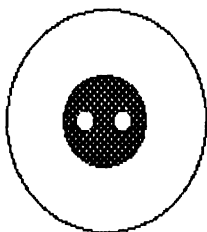
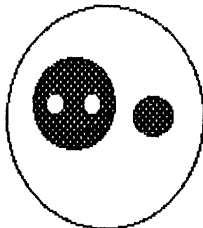
Here is a class of real $K3$ -surfaces with a condition on its Picard lattice, which is especially remarkable because of its connection with 4-dimensional Riemannian geometry. The surfaces from this class are said to be *K3R-surfaces*. Its lattice S is given by

$$S = [h; e_1, e_2, e_3, e_4; f_1, f_2, f_3, f_4; e = \frac{1}{2}(h - e_1 - e_2 - e_3 - e_4), \\ f = \frac{1}{2}(h - f_1 - f_2 - f_3 - f_4)],$$

where the elements h, e_i , and f_i are orthogonal to each other, $h^2 = 8$, $e_i^2 = f_i^2 = -2$, $f^2 = e^2 = 0$. The polarization is given by the vector $h_1 = e + f = h - \frac{1}{2}(e_1 + \dots + e_4 + f_1 + \dots + f_4)$, $h_1^2 = 4$ (note that we have changed slightly our previous notation). For simplicity we will restrict ourselves to the case of pseudo-Riemannian manifolds with Lorentz metric (following [42, 46]). In this case the action σ is given by $\sigma^*(h) = -h$, $\sigma^*(e_i) = -f_i$, $\sigma^*(f_i) = -e_i$.

Let \mathcal{V} be a 4-dimensional pseudo-Riemannian manifold with a Lorentz metric, $G = G_x$ be the quadratic form of the metric on the tangent space $\mathcal{T} = \mathcal{T}_x$ at a point x of \mathcal{V} . Let $R = R_x$ be the quadratic form of the Riemannian curvature in $\Lambda^2 \mathcal{T}$. Let $\mathcal{G} \subset \mathbf{P}^3 = \mathbf{P}(\mathcal{T})$ be the quadric with equation $G = 0$, $\text{Gr}(2, 4) \subset \mathbf{P}^5 = \mathbf{P}(\Lambda^2 \mathcal{T})$ be the Grassmann variety of lines in \mathbf{P}^3 (all objects are defined over \mathbf{R}). The algebraic variety

$$X = \{(q, l) \in \mathbf{P}^3 \times \mathbf{P}^5 : q \in \mathcal{G}, l \in \text{Gr}(2, 4), R(l) = 0 \text{ and } l \text{ touches } \mathcal{G} \text{ at } q\} \\ \subset \mathbf{P}^3 \times \mathbf{P}^5$$

Case $0_k, 0 \leq k \leq 4$ Case $1_k, 0 \leq k \leq 4$ Case $2_k, k = 0, 1$  $k = 0$  $k = 1$ Case $k, 3 \leq k \leq 5$ 

Case 1.1

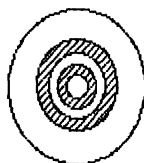


Figure 4

is defined uniquely up to automorphisms of \mathbf{P}^3 and \mathbf{P}^5 and is said to be the *Tyurin invariant* of the point $x \in \mathcal{V}$ (see [53, 54] and also [46]). Let $p_1: X \rightarrow \mathbf{P}^3$ and $p: X \rightarrow \mathbf{P}^5$ be the projections. For sufficiently general metric at a point $x \in \mathcal{V}$, the variety X is a nonsingular $K3$ -surface and $p_i: X \rightarrow \mathcal{G} \subset \mathbf{P}^3$ is a double covering of \mathcal{G} which is ramified over a nonsingular curve of bidegree $(4, 4)$. In this case the Tyurin invariants are said to be strongly nonsingular. For them the curve C has two tangency points with 4 lines $\tilde{E}_1, \dots, \tilde{E}_4$ from one ruling of \mathcal{G} , and 4 lines $\tilde{F}_1, \dots, \tilde{F}_4$ from the other one. The divisor classes e_i and f_i , respectively, are the classes of nonsingular rational curves $E_i = \{(x, \tilde{E}_i): x \in \tilde{E}_i\}$ and $F_j = \{(x, \tilde{F}_j): x \in \tilde{F}_j\}$ on X . The divisor classes h_1 and h , respectively, are the inverse images of hyperplane sections under the projections p_1 and p .

The coarse projective classification of Tyurin invariants allows us to separate 4-dimensional Lorentz manifolds according to the types of their metric. It is analogous to distinguishing 2-dimensional Riemannian manifolds by the positivity or the negativity of their curvature. Moreover, this partition is nontrivial under the usual positivity assumptions of the energy-momentum tensor of the

matter in general relativity theory. Hence, it is plausible that this partition has physical sense, separating points of our space-time into different types.

At the present time only the first part (a) of the coarse projective classification of Tyurin invariants is solved. There are exactly 20 possibilities for the isomorphism classes of the triplets (L, σ^*, S) corresponding to the Tyurin invariants [42, 46]. Each triplet defines a configuration of the image $p_1(X(\mathbf{R})) \subset \mathcal{G}(\mathbf{R}) = S^2$, where S^2 is the 2-dimensional light sphere. In other words, it defines the isotopy type of the Tyurin invariant (which obviously gives the topology of $X(\mathbf{R})$). These configurations are presented in Figure 4, where the image $p_1(X(\mathbf{R}))$ is shaded. The configuration 1_0 gives 3 triplets, each of the configurations 0_0 and 0_4 gives 2 triplets, and each other configuration gives one triplet.

The second part (b) of this classification is not known yet. However, there are some interesting and visual invariants of this classification (see [46]). For example, the lines tangent to the light sphere S^2 at unshaded points of Figure 4 have definite sign of the curvature R . In such a way, the unshaded pieces of S^2 have signs $+$ or $-$. Their possible distributions are yet unknown.

BIBLIOGRAPHY

1. E. M. Andreev, *On convex polyhedra in Lobachevsky spaces*, Mat. Sb. (N.S.) **81** (1970), 445–478 = Math. USSR-Sb. **10** (1970).
2. —, *On convex polyhedra of finite volume in Lobachevsky spaces*, Mat. Sb. (N.S.) **83** (1970), 256–260 = Math. USSR-Sb. **12** (1970).
3. W. Barth and C. Peters, *Automorphisms of Enriques surfaces*, Invent. Math. **73** (1983), 383–411.
4. N. Bourbaki, *Groupes et algèbres de Lie*, Chaps. 4–6, Hermann, Paris, 1968.
5. A. Coble, *The ten nodes of the rational sextic and the Cayley symmetroid*, Amer. J. Math. **41** (1919), 243–265.
6. J. H. Conway, *The automorphism group of the 26-dimensional even unimodular Lorentzian lattice*, J. Algebra **80** (1983), 159–163.
7. F. Cossec and I. Dolgachev, *On automorphisms of nodal Enriques surfaces*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), 247–249.
8. H. S. M. Coxeter, *Discrete groups generated by reflections*, Ann. Math. **35** (1934), 588–621.
9. V. I. Danilov, *The geometry of toric varieties*, Uspekhi Mat. Nauk **33** (1978), 85–134 = Russian Math. Surveys **33** (1978), 97–154.
10. M. Demazure, *Surfaces de Del Pezzo*, Lecture Notes in Math., vol. 777, Springer-Verlag, 1980, pp. 23–69.
11. I. Dolgachev, *Integral quadratic forms: applications to algebraic geometry (after V. Nikulin)*, Seminar Bourbaki 1982/83, Astérisque, 105–106, Soc. Math. France, Paris, 1984, pp. 251–278.
12. —, *On automorphisms of Enriques surfaces*, Invent. Math. **76** (1984), 163–177.
13. P. Du Val, *On the Kantor group of a set of points in a plane*, Proc. London Math. Soc. **42** (1937), 18–51.
14. W. Dyck, *Gruppentheoretische Studien*, Math. Ann. **20** (1882), 1–45.
15. G. Fano, *Superficie algebriche di genere zero e bigenero uno e loro casi particolari*, Rend. Circ. Mat. Palermo **29** (1910), 98–118.
16. B. Harbourne, *Complete linear systems on rational surfaces*, Trans. Amer. Math. Soc. **289** (1985), 213–226.
17. —, *Blowing-up of P^2 and their blowing down*, Duke Math. J. **52** (1985), 129–148.
18. —, *Rational surfaces with infinite automorphism group and no antipluricanonical curve*, Preprint, 1985.

19. —, *Automorphisms of rational surfaces*, Preprint, 1985.
20. F. Hirzebruch, *Arrangements of lines and algebraic surfaces*, Arithmetic and Geometry, Vol. II, Progr. Math. 36, Birkhauser, Boston, Mass., 1983, pp. 113–140.
21. V. M. Kharlamov, *Topological types of nonsingular surfaces of degree 4 in RP^3* , Funktsional. Anal. i Prilozhen. **10** (1976), 55–68 = Functional Anal. Appl. **10** (1976).
22. —, *Isotopy types of nonsingular surfaces of degree 4 in RP^3* , Funktsional. Anal. i Prilozhen. **12** (1978), 86–87 = Functional Anal. Appl. **12** (1978).
23. —, *Real algebraic surfaces*, Proc. Internat. Congr. Math. (Helsinki, 1978), Acad. Sci., Fennica, Helsinki, 1980, pp. 421–428.
24. —, *On the classification of nonsingular surfaces of degree 4 in RP^3 with respect to rigid isotopies*, Funktsional. Anal. i Prilozhen. **18** (1984), 49–56 = Functional Anal. Appl. **18** (1984).
25. A. G. Khovanskii, *Linear programming and a geometry of convex polyhedra*, Proc. VNI Systems Researches 7, Moscow, 1985, pp. 73–81. (Russian)
26. —, *Linear programming and a generalization of Nikulin's theorem*, Proc. VNI Systems Researches 7, Moscow, 1985, pp. 81–87. (Russian)
27. —, *Polyhedron hyperplane sections, torical varieties and discrete groups in Lobachevsky space*, Funktsional. Anal. i Prilozhen. **20** (1986), 50–61 = Functional Anal. Appl. **20** (1986).
28. S. Kondō, *Enriques surfaces with finite automorphism groups*, Preprint, 1985.
29. Vik. S. Kulikov, *Degenerations of K3-surfaces and Enriques surfaces*, Izv. Akad. Nauk SSSR Ser. Mat. **41** (1977), 1008–1042 = Math. USSR-Izv. **11** (1977), 957–989.
30. E. Looijenga, *The discriminant of a real simple singularity*, Compositio Math. **37** (1978), 51–62.
31. —, *Rational surfaces with effective anticanonical divisor*, Ann. Math. **114** (1981), 267–322.
32. V. S. Makarov, *On a certain class of discrete Lobachevsky space groups with infinite fundamental domain of finite measure*, Dokl. Akad. Nauk SSSR **167** (1966), 30–33 = Soviet Math. Dokl. **7** (1966).
33. —, *Geometric methods for construction of discrete motion groups in Lobachevsky space*, Problems of Geometry, vol. 15, Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Informatsii, Moscow, 1983, pp. 3–59 = J. Soviet Math. (to appear).
34. Ju. I. Manin, *Cubic forms: algebra, geometry, arithmetic*, "Nauka", Moscow, 1972; English transl., North-Holland, Amsterdam, 1973 and American Elsevier, New York, 1974.
35. M. Nagata, *On rational surfaces*. I, II, Mem. Coll. Sci. Kyoto (A) **32** (1960), 351–370; **33** (1960), 271–293.
36. Y. Namikawa, *Periods of Enriques surfaces*, Math. Ann. **270** (1985), 201–222.
37. V. V. Nikulin, *Integral symmetric bilinear forms and some of their geometric applications*, Izv. Akad. Nauk SSSR Ser. Mat. **43** (1979), 103–167 = Math. USSR-Izv. **14** (1980), 103–167.
38. —, *On factor groups of the automorphism groups of hyperbolic forms modulo subgroups generated by 2-reflections*, Dokl. Akad. Nauk SSSR **248** (1979), 1307–1309 = Soviet Math. Dokl. **20** (1979), 1156–1158.
39. —, *On the quotient groups of the automorphism groups of hyperbolic forms by the subgroups generated by 2-reflections, Algebraic-geometric applications*, Current Problems in Math., vol. 18, Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Informatsii, Moscow, 1981, pp. 3–114 = J. Soviet Math. **22** (1983), 1401–1476.
40. —, *On arithmetic groups generated by reflections in Lobachevsky spaces*, Izv. Akad. Nauk SSSR Ser. Mat. **44** (1980), 637–669 = Math. USSR-Izv. **16** (1981), 537–601.
41. —, *On classification of arithmetic groups generated by reflections in Lobachevsky spaces*, Izv. Akad. Nauk SSSR Ser. Mat. **45** (1981), 113–142 = Math. USSR-Izv. **18** (1982), 99–123.
42. —, *Involutions of integral quadratic forms and their application to algebraic geometry*, Izv. Akad. Nauk SSSR Ser. Mat. **47** (1983), 109–188 = Math. USSR-Izv. (to appear).
43. —, *Surfaces of type K3 with finite automorphism group and Picard group of rank three*, Proc. Steklov Inst. Math. **165** (1984), 119–142 = Trudy Inst. Steklov. **3** (1985).

44. —, *On a description of the automorphism groups of an Enriques surfaces*, Dokl. Akad. Nauk SSSR **277** (1984), 1324–1327 = Soviet Math. Dokl. **30** (1984), 282–285.
45. —, *Filtrations of 2-elementary forms and involutions of integral bilinear symmetric and skew symmetric forms*, Izv. Akad. Nauk SSSR Ser. Math. **49** (1985), 847–873 = Math. USSR-Izv. (to appear).
46. —, *Local invariants of 4-dimensional pseudo-Riemannian manifolds with Lorentz metric*, Problems of Geometry, vol. 17, Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Informatsii, Moscow, 1985, pp. 87–130.
47. I. I. Piatetsky-Shapiro and I. R. Šafarevich, *A Torelli theorem for algebraic surfaces of type K3*, Izv. Akad. Nauk SSSR Ser. Mat. **35** (1971), 530–572 = Math. USSR-Izv. **5** (1971), 547–588.
48. H. Poincaré, *Theorie des groupes fuchsienues*, Acta Math. **1** (1882), 1–62.
49. M. N. Prokhorov, *Absence of discrete reflection groups with a noncompact fundamental polyhedron of finite volume in Lobachevsky spaces of large dimension*, Izv. Akad. Nauk SSSR Ser. Mat. **50** (1986), 413–424 = Math. USSR-Izv. (to appear).
50. F. Severi, *Complementi alla theoria della base per la totalita delle curve di una superficie algebrica*, Rend. Circ. Mat. Palermo **39** (1910), 265–288.
51. O. V. Švartsman, *Discrete reflection groups in a complex ball*, Funktsional. Anal. i Prilozhen. **18** (1984), 88–89 = Functional Anal. Appl. **18** (1984).
52. R. Stanley, *The upper bound conjecture and Cohen-Macaulay rings*, Stud. Appl. Math. **54** (1975), 135–142.
53. A. N. Tyurin, *A local invariant of a Riemannian manifold*, Izv. Akad. Nauk SSSR Ser. Mat. **45** (1981), 824–851 = Math. USSR-Izv. **19** (1982).
54. —, *Local and global invariants of a four dimensional pseudo-Riemannian manifolds*, Proc. Steklov Inst. Math. **165** (1984), 205–219 = Trudy Mat. Inst. Steklov. **3** (1985).
55. E. B. Vinberg, *Discrete groups generated by reflections in Lobachevsky spaces*, Math. Sb. (N.S.) **72** (1967), 471–488 = Math. USSR-Sb. **1** (1967), 429–444.
56. —, *On unimodular integral quadratic forms*, Funktsional. Anal. i Prilozhen. **6** (1972), 24–31 = Functional Anal. Appl. **6** (1972), 105–111.
57. —, *On the group of unit elements of certain quadratic forms*, Mat. Sb. (N.S.) **87** (1972), 18–36 = Math. USSR-Sb. **16** (1972), 17–35.
58. —, *Some arithmetic discrete groups in Lobachevsky spaces*, Discrete Subgroups of Lie Groups and Applications to Moduli, Oxford Univ. Press, Bombay, 1975, pp. 323–348.
59. —, *Absence of crystallographic reflection groups in Lobachevsky spaces of large dimension*, Funktsional. Anal. i Prilozhen. **15** (1981), 67–68 = Functional Anal. Appl. **15** (1981), 128–130.
60. —, *The two most algebraic K3 surfaces*, Math. Ann. **265** (1983), 1–21.
61. —, *Discrete reflection groups in Lobachevsky spaces*, Proc. Internat. Congr. Math. (Warsaw, 1983), pp. 593–601.
62. —, *Absence of crystallographic reflection groups in Lobachevsky spaces of large dimension*, Trudy Moskov. Mat. Obschch. **47** (1984), 68–102 = Trans. Moscow Math. Soc. (to appear).
63. —, *Hyperbolic reflection groups*, Uspekhi Mat. Nauk **40** (1985), 29–66 = Russian Math. Surveys (to appear).
64. E. B. Vinberg and I. M. Kaplinskaya, *On the groups $O_{18,1}(\mathbb{Z})$ and $O_{19,1}(\mathbb{Z})$* , Dokl. Akad. Nauk SSSR **238** (1978), 1273–1275 = Soviet Math. Dokl. **19** (1978), 194–197.
65. O. J. Viro, *Progress in the last 5 years in topology of real algebraic varieties*, Proc. Internat. Congr. Math. (Warsaw, 1983), pp. 603–619. (Russian)

Numerical Geometry of Algebraic Varieties

V. V. SHOKUROV

This report is devoted to the achievements and difficulties in the program of constructing minimal models for algebraic varieties of any dimension. Two essential results have been obtained within the framework of this program: the contraction and the cone theorems. Of no less importance are the working out of the terminology and the statement of two key conjectures: the flip and the termination conjectures, whose solutions would lead to the completion of the whole program. These conjectures have been proved in some special cases; for example, they both hold for toric morphisms and the latter holds in dimension ≤ 4 . The final results are in many cases due to many mathematicians from all over the world, their names following the title of a result in the order of their contribution to it.

1. \mathbf{Q} -divisors. Let X be a normal projective algebraic variety over the complex number field. A \mathbf{Q} -Cartier divisor on X , or simply a \mathbf{Q} -divisor on X , is an element $D \in \text{Div}_{\mathbf{Q}} X = \text{Div } X \otimes \mathbf{Q}$, where $\text{Div } X$ is the group of Cartier divisors on X ; in other words, a \mathbf{Q} -divisor is a linear combination of Cartier divisors with coefficients in \mathbf{Q} . The group $\text{Div}_{\mathbf{Q}} X$ also contains certain Weil divisors of X , namely, such D that $rD \in \text{Div } X$ for some $0 \neq r \in \mathbf{Z}$; for then $D = (1/r)(rD)$. A variety X is said to be \mathbf{Q} -factorial if every Weil divisor of X is \mathbf{Q} -Cartier in this sense.

A *canonical divisor* K_X is a Weil divisor on X such that

$$\mathcal{O}_{X_{\text{reg}}}(K_X) = \Omega_{X_{\text{reg}}}^{\dim X},$$

where X_{reg} is the nonsingular locus of X . Since $\text{codim}(X - X_{\text{reg}}) \geq 2$, a canonical divisor K_X is well defined and $K_X = K_{X_{\text{reg}}}$ under natural identification of Weil divisors on X and on X_{reg} . In general a canonical divisor K_X is not a \mathbf{Q} -divisor. But when K_X is a \mathbf{Q} -divisor a variety X is said to be \mathbf{Q} -Gorenstein. A minimal number r such that $rK_X \in \text{Div } X$ is called the *index* of a \mathbf{Q} -Gorenstein variety X . These definitions do not depend on the choice of K_X because two canonical divisors differ by a Cartier divisor. Obviously, every \mathbf{Q} -factorial variety X is \mathbf{Q} -Gorenstein.

2. Intersections. Any morphism $f: Y \rightarrow X$ of normal projective varieties induces the inverse image map $f^*: \operatorname{Div}_{\mathbf{Q}} X \rightarrow \operatorname{Div}_{\mathbf{Q}} Y$. This gives a natural intersection theory for \mathbf{Q} -divisors. For example, we have a bilinear pairing

$$(\cdot): \operatorname{Div}_{\mathbf{Q}} X \times Z_1 X \rightarrow \mathbf{Q},$$

where $Z_1 X$ is the free Abelian group generated by curves on X .

A \mathbf{Q} -divisor $D \in \operatorname{Div}_{\mathbf{Q}} X$ is said to be *numerically effective* or briefly *nef* if $(D.C) \geq 0$ for every curve $C \subseteq X$. To each nef \mathbf{Q} -divisor D we associate its numerical dimension

$$\nu(D) = \max\{k | D^k \neq 0\},$$

where $D^k \neq 0$ means that $D^k C \neq 0$ for some k -cycle C on X . Obviously,

$$\max\{0, \kappa(D)\} \leq \nu(D) \leq n = \dim X,$$

where $\kappa(D)$ is the Iitaka D -dimension of X .

Let D be a nef \mathbf{Q} -divisor. Then the following conditions are equivalent:

- 2.1. $\nu(D) = n$;
- 2.2. $D^n \stackrel{\text{df}}{=} D^n X > 0$;
- 2.3. $\kappa(D) = n$.

A nef \mathbf{Q} -divisor D satisfying any one of these conditions is said to be *big*.

3. Singularities. Throughout what follows we assume that X is \mathbf{Q} -Gorenstein. Let $f: Y \rightarrow X$ be a desingularization of X . Then for any canonical divisor K_X there exists a canonical divisor K_Y such that

$$K_Y = f^* K_X + \sum a_i E_i,$$

where $a_i \in \mathbf{Q}$ and the E_i vary over all prime divisors which are exceptional with respect to f . The divisor $\sum a_i E_i$ in the above equation does not depend on the choice of canonical divisors K_X and K_Y . The coefficient a_i is referred to as the *discrepancy* at E_i . A variety X is said to have only *terminal* (resp. *canonical*) singularities if all discrepancies $a_i > 0$ (resp. $a_i \geq 0$). This notion does not depend on the choice of a desingularization of X .

3.1. THEOREM. *All canonical and terminal singularities are rational.*

Note that when $\dim X = 2$, X has only terminal singularities if and only if X is nonsingular. Surface canonical singularities have many different descriptions and are also known as du Val singularities.

4. Models. A variety X with only canonical singularities is said to be *canonical* if K_X is ample. A canonical variety, birationally equivalent to a given variety Y , is a *canonical model* of Y . It follows from the invariance of the Kodaira dimension $\kappa(X) \stackrel{\text{df}}{=} \kappa(K_X)$ for birational transformations of nonsingular varieties and from the definition of canonical singularities that a nonsingular variety X has a canonical model X_{can} only if $\kappa(X) = \dim X$, i.e., X is of *general type*.

4.1. THEOREM (REID). *A nonsingular variety X of general type has a canonical model X_{can} if and only if the canonical ring of X*

$$R(X) = \bigoplus_{m \geq 0} H^0(X, \mathcal{O}_X(mK_X))$$

is finitely generated. Moreover,

$$X_{\text{can}} = \text{Proj } R(X).$$

4.2. COROLLARY. *Any nonsingular variety X of general type has a unique canonical model if it exists.*

A variety X with only terminal singularities is said to be *minimal* if K_X is nef. A minimal variety, birationally equivalent to a given variety Y , is a *minimal model* of Y . A nonsingular algebraic surface X has a minimal model if and only if $\kappa(X) \geq 0$ and it is always unique. But when $\dim X \geq 3$ there may be many minimal models of X and there is certainly no absolute minimal model in general. In the 3-fold case, key counterexamples concern (-2) -curves and the Kulikov modification or more generally Pagoda modifications in these curves. However, there is hope that at least in dimension three any nonsingular variety of general type has only finitely many minimal models.

A morphism $h: X \rightarrow Z$ of normal projective varieties is said to be a *relative anticanonical variety* if

4.3. $h_* \mathcal{O}_X = \mathcal{O}_Z$ and

4.4. $-K_X$ is relatively ample for h , i.e., $\mathcal{O}_X(-rK_X)|_Y$ is an ample invertible sheaf on any fibre $Y = h^{-1}(z)$, $z \in Z$.

In addition, we assume that X has only canonical singularities and $\dim Y \geq 1$ for all fibres Y . If Z is a point, an anticanonical variety X is said to be a *Fano variety*.

A relative anticanonical variety $h: X \rightarrow Z$ has the following properties:

4.5. $R^i h_* \mathcal{O}_X = 0$ for all $i > 0$ and, in particular, $\chi(\mathcal{O}_X) = \chi(\mathcal{O}_Z)$.

4.6. (Kollár). Z has only rational singularities.

4.7. The general fibre $Y = h^{-1}(z)$, $z \in Z$, is a Fano variety with only canonical singularities. Moreover, $H^i(Y, \mathcal{O}_Y) = 0$ for all $i > 0$ and the group scheme $\text{Pic } Y$ is discrete and torsion free. If X has only terminal singularities, then so does Y .

Note that \mathbf{P}^1 is a unique Fano variety in dimension one. So the general fibre of a relative anticanonical variety with $\dim Z = \dim X - 1$ is \mathbf{P}^1 , and this variety is also known as *conic bundle*. Two-dimensional Fano varieties are usually called Del Pezzo surfaces. Respectively, a relative anticanonical variety with $\dim Z = \dim X - 2$ is a *Del Pezzo fibre space*.

A relative anticanonical variety $h: X \rightarrow Z$ with only terminal singularities on X such that X is birationally equivalent to a given variety Y is called a *relative anticanonical model* of Y . Obviously, Y has a relative anticanonical model only if $\kappa(Y) = -\infty$. As a rule, a relative anticanonical model is not unique. For example, 3-folds have three possible types of relative anticanonical models: conic bundles over surfaces, Del Pezzo fibre spaces over curves and Fano

3-folds. \mathbf{P}^3 has infinitely many relative anticanonical models of every type and all of the types may be on the same birational transform. Indeed, $\mathbf{P}^1 \times \mathbf{P}^2 \rightarrow \mathbf{P}^2$ is a conic bundle, $\mathbf{P}^1 \times \mathbf{P}^2 \rightarrow \mathbf{P}^1$ is a Del Pezzo fibre space, and $\mathbf{P}^1 \times \mathbf{P}^2$ is a Fano 3-fold. But a nonsingular quartic 3-fold is a Fano 3-fold, and it is its only relative anticanonical model known to date. Note that a nonsingular algebraic surface X has a relative anticanonical model if and only if $\kappa(X) = -\infty$.

The following conjecture is a main driving force and at the same time is the stumbling-block in the development of a modern algebraic geometry in higher dimensions.

MODEL CONJECTURE. Every nonsingular projective (or complete) algebraic variety has a minimal model or a relative anticanonical model.

In dimension two this conjecture is true due to the Enriques classification of algebraic surfaces. Now we turn to successes and difficulties met in carrying out the minimal model program.

5. Contractions. By a contraction we mean a surjective morphism $\varphi: X \rightarrow Z$ of normal projective varieties which satisfies property 4.3. Every such morphism is a morphism $\varphi|_{|mD|}: X \rightarrow \varphi|_{|mD|}(X) \subseteq \mathbf{P}^N$, given by a complete linear system $|mD|$ for some $m > 0$, where $D = \varphi^*H$ and H is an ample divisor on $Z = \varphi|_{|mD|}(X)$. Note that the linear system $|mD|$ is *free*, i.e., $|mD|$ has no base points, and hence the divisor D is *semiample*, i.e., $|mD|$ is free for some $m > 0$. Conversely, any semiample divisor D on X defines a contraction $\varphi: X \rightarrow Z$ such that $mD \sim \varphi^*H$ for some $m > 0$, where H is a (very) ample divisor on Z and \sim means the linear equivalence of divisors. The following statement gives a numerical criterion for semiampleness.

CONTRACTION THEOREM (KAWAMATA, REID, SHOKUROV). *Let D be a Cartier divisor on a variety X with only canonical singularities such that*

5.1. D is nef.

5.2. $aD - K_X$ is nef and big for some $a > 0$. Then D is semiample. Moreover, D is stably free, that is, for all $m \gg 0$, the linear system $|mD|$ is free. Equivalently there exists a contraction $\varphi: X \rightarrow Z$ such that $D \sim \varphi^*H$ for an ample divisor $H \in \text{Div } Z$.

The contraction φ has the following properties:

5.3. $R^i\varphi_*\mathcal{O}_X = 0$ for all $i > 0$ and, in particular, $\chi(\mathcal{O}_X) = \chi(\mathcal{O}_Z)$ (cf. 4.5).

5.4. There exists a unique decomposition $\varphi = h \circ g$ of φ as a composite of contractions

$$X \xrightarrow{g} X^- \xrightarrow{h} Z,$$

where X^- is a normal projective variety with only canonical singularities, such that g is a birational morphism with $K_X = g^*K_{X^-}$, and $-K_{X^-}$ is relatively ample for h (cf. 4.4). Moreover, h is a relative anticanonical variety when φ is not birational, i.e., $\dim Z < \dim X$.

The proof of the contraction theorem uses the vanishing and nonvanishing theorems. Applying the contraction theorem to a divisor $D = rK_X$ on a minimal variety X with big K_X we obtain

5.5. COROLLARY (KAWAMATA, BENVENISTE, SHOKUROV). *If a non-singular projective variety Y of general type has a minimal model X , then there exists a canonical model Y_{can} and a birational contraction $\varphi: X \rightarrow Y_{\text{can}}$ such that $\varphi^*K_{Y_{\text{can}}} = K_X$.*

6. Kleiman-Mori cone. Two 1-cycles $z_1, z_2 \in Z_1X$ are said to be *numerically equivalent* if, for any divisor $D \in \text{Div } X$, $(D.z_1) = (D.z_2)$, and we then write $z_1 \equiv z_2$. Let $N_1X = (Z_1X / \equiv) \otimes \mathbf{R}$. This is a finite-dimensional real linear space and $\rho(X) \stackrel{\text{df}}{=} \dim_{\mathbf{R}} N_1X$ is known as the *Picard number* of X .

Let V be a finite-dimensional real linear space. A convex subset S in V is said to be polyhedral if

$$S = \{v \in V \mid f_1(v) \leq 0, \dots, f_N(v) \leq 0\},$$

where f_1, \dots, f_N are linear forms on V . If $V = W \otimes \mathbf{R}$, where W is a maximal lattice in V and f_1, \dots, f_N are integral on W , then a subset S is said to be *rational polyhedral*. A subset $S \subseteq V$ is said to be *locally polyhedral* (resp. rational polyhedral) if it is local like a polyhedral (resp. rational polyhedral) set in V .

By the *Kleiman-Mori cone* we mean the closure $\overline{NE}(X)$ of the convex cone $NE(X)$ generated by effective 1-cycles or, equivalently, by curves in N_1X .

CONE THEOREM (MORI, KAWAMATA, SHOKUROV, REID, KOLLÁR). *Let X be a projective variety with only canonical singularities. Then the "half-cone" $\overline{NE}_-(X) \stackrel{\text{df}}{=} \{z \in \overline{NE}(X) \mid (K_X.z) < 0\}$ is locally rational polyhedral.*

Note that the cone theorem also holds for nonsingular projective varieties over an arbitrary algebraically closed field $[\mathbf{M}]$.

A face F of the cone $NE(X)$ is said to be *extremal* if $F - \{0\} \subseteq \overline{NE}_-(X)$. An *extremal ray* is a one-dimensional face.

6.1. COROLLARY (MORI). *If K_X is not nef, then there exists a nontrivial extremal face and, hence, an extremal ray.*

7. Extremal contractions. Let X be a projective variety with only terminal singularities; for example, X may be nonsingular. If K_X is nef, then X is a minimal variety. Otherwise, there is a nontrivial extremal face F of $\overline{NE}(X)$. Then F uniquely determines a contraction $\varphi = \text{cont}_F: X \rightarrow Z$, such that

7.1. For any curve $C \subseteq X$, we have $\varphi(C) = \text{pt.} \in Z \Leftrightarrow$ the numerical class of C belongs to F . cont_F is called the *contraction of the face F* . Indeed, due to the rationality in the cone theorem we have a nef divisor $D \in \text{Div } X$ such that

$$\mathbf{7.2.} \quad F = \{z \in \overline{NE}(X) \mid (D.z) = 0\}.$$

Moreover, D satisfies condition 5.2 in the contraction theorem and φ is determined by D . Obviously, φ does not depend on the choice of such D . Note also

that there exists a curve $C \subseteq X$ contracted by φ because D is not ample in the sense of Kleiman [K1]. If $\dim Z < \dim X$, then $\varphi: X \rightarrow Z$ is a relative anticanonical variety. Such extremal contractions are of *fibre type*.

For simplicity throughout what follows we assume in addition that X is \mathbf{Q} -factorial and $F = R$ is an extremal ray. Then $\varphi = \text{cont}_R: X \rightarrow Z$ has the following properties and classification:

7.3. There is an exact sequence

$$0 \rightarrow \mathbf{R}[C] \rightarrow N_1 X \rightarrow N_1 Z \rightarrow 0,$$

where C is a curve with the numerical class $[C] \in R$. Hence $\rho(Z) = \rho(X) - 1$.

7.4. If $\dim Z < \dim X$, then φ is of fibre type.

7.5. If $\dim Z = \dim X$ and there exists a prime Weil divisor E on X such that $\dim \varphi(E) < \dim E$, then Z is again a \mathbf{Q} -factorial variety with only terminal singularities. Such a divisor E is unique and $E = \bigcup C$, where the C vary over all curves with $[C] \in R$. This contraction φ is said to be of *divisorial type*, and the divisor E is said to be the *exceptional divisor associated to the ray R*.

7.6. If $\dim Z = \dim X$ and φ is not of divisorial type, then there exists a subvariety $E \subset X$ such that $1 \leq \dim E \leq \dim X - 2$ and $\varphi|_{X-E}$ is an isomorphism. Such a contraction φ is said to be of *flipping type*, and a minimal subvariety E with the above property is said to be the *exceptional locus* of φ .

The major difficulty in carrying out the minimal model program in dimension ≥ 3 arises when φ is of flipping type because in this case Z is not \mathbf{Q} -factorial; still worse, it is not \mathbf{Q} -Gorenstein. If $\dim X = 2$ or X is a nonsingular 3-fold [M] we have no *flipping* extremal rays, i.e., rays with flipping contractions.

8. Extremal modifications. There is an entirely natural

FLIP CONJECTURE. Any extremal ray R of flipping type has an *adjoint diagram* or a *flip*. This should be a commutative diagram

$$\begin{array}{ccc} X & \xrightarrow{\text{tr}_R} & X^+ \\ \varphi = \text{cont}_R \searrow & & \swarrow \varphi^+ \\ & Z & \end{array}$$

consisting of:

8.1. A normal projective \mathbf{Q} -Gorenstein variety X^+ , which is said to be an *extremal transform* of X in the ray R ,

8.2. A rational map $\text{tr}_R: X \rightarrow X^+$ which is an isomorphism except for loci of codimension ≥ 2 . The map tr_R is an *extremal modification* in the ray R .

8.3. A contraction φ^+ such that a canonical divisor K_{X^+} is relatively ample for φ^+ . (Recall that $-K_X$ is relatively ample for φ .)

8.4. PROPOSITION (on a platter on one's head). *An adjoint diagram exists if and only if for some ample divisor $H \in \text{Div } Z$ and any integer $m \gg 0$ the adjoint-canonical ring*

$$R(m\varphi^*H + K_X) = \bigoplus_{n \geq 0} H^0(X, \mathcal{O}_X(n(m\varphi^*H + K_X)))$$

is finitely generated. Moreover, in this case

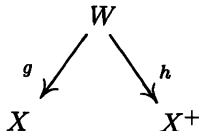
$$X^+ = \text{Proj } R(m\varphi^*H + K_X).$$

A flip has the following properties:

8.5. If an extremal ray has a flip, then it is unique.

8.6. X^+ is \mathbf{Q} -factorial and has only terminal singularities.

8.7. Moreover, for any common desingularization



we have $a_i^+ \geq a_i$, with $a_i^+ > a_i$, if and only if $\varphi \circ g(E_i) \subseteq E$, where a_i and a_i^+ are discrepancies of g and h at E_i respectively and E is the exceptional locus of φ .

8.8. $\rho(X^+) = \rho(X)$.

The flip conjecture holds for flipping extremal rays with toric contractions, so there is good reason to believe the flip conjecture to hold in general. In dimension three we have

9. Some sufficient conditions for the existence of a flip. The best one now is

KAWAMATA CONDITION. Let $\varphi: X \rightarrow Z$ be a contraction for a flipping extremal ray R and $\dim X = 3$. Then a flip in R exists if the divisor $K_X + D/2$ is log-terminal for some effective reduced Weil divisor $D \in |-2K_X + \varphi^*H|$, where $H \in \text{Div } Z$.

Note that D is \mathbf{Q} -Cartier. A divisor $K_X + D/2$ is said to be *log-terminal* if there is a desingularization $f: Y \rightarrow X$ for which

(i) $D' + \sum E_i$ is a normal crossing divisor, where D' is the strict transform of D by f and $\sum E_i$ is the exceptional locus of f ;

(ii) $2(a_i + 1) > r_i$ for all i , where a_i is the discrepancy at E_i and $f^*D = D' + \sum r_i E_i$.

It is known that (ii) does not depend on the choice of desingularization with (i). So the Kawamata condition states that a flip exists if a general divisor $D \in |-2K_X + \varphi^*H|$ has mild singularities in the above sense.

Another condition concerns semistable extremal rays. A reduced Weil divisor $D \subset X$ is said to be *simple* if D is \mathbf{Q} -Cartier and there is a desingularization $f: Y \rightarrow X$ such that $f^*D = D' + \sum_{i=1}^d E_i$ and it is a divisor with normal crossing, where D' is the strict transform of D by f and $\sum_{i=1}^d E_i$ is the exceptional locus of f . A minimal d for such desingularizations is the *depth* of D . An extremal ray R is semistable when there exists a simple divisor D on X such that any curve $C \subseteq D$ provided the numerical class of C belongs to R .

9.1. THEOREM (TSUNODA, SHOKUROV, MORI, KAWAMATA). *If R is a flipping semistable ray, then there exists a flip in R . Besides, if R is semistable for D , then $\text{tr}_R D$ is simple and the depth of $\text{tr}_R D$ is less than that of D .*

We may apply the last theorem to a semistable family $f: X \rightarrow C$ of surfaces over a curve if the general fibre $f^{-1}(c)$, $c \in C$, is a minimal surface. For example, in such a way a Kulikov relative model of a stable one-dimensional family of K3 surfaces is obtained after some easy modification [S1].

10. Termination. Now we turn to yet another difficulty in the minimal model program. First note that $\rho(Z) = \rho(X) - 1$ for divisorial contractions of extremal rays and $\rho(X^+) = \rho(X)$ for modifications in flipping extremal rays. So there arises the problem as to how long we may produce modifications in flipping extremal rays. The

TERMINATION CONJECTURE states that a sequence of modifications in flipping extremal rays must always *terminate*, i.e., there does not exist an infinite sequence of such modifications.

This conjecture together with the flip conjecture implies that there exists a finite chain of flipping modifications

$$X \dashrightarrow X^+ \dashrightarrow \cdots \dashrightarrow X^{(+n)} = Z$$

such that either

- (i) K_Z is nef, or
- (ii) Z has a contraction of fibre type, or
- (iii) Z has a contraction of divisorial type.

So they imply the model conjecture.

10.1. THEOREM (SHOKUROV, KAWAMATA). *The termination conjecture holds when $\dim X \leq 4$.*

10.2. COROLLARY. *The flip conjecture implies the model conjecture in dimension ≤ 4 .*

11. Abundance. Let X be a projective normal variety and $\text{Div}_{\mathbf{Q}} X$ be a nef \mathbf{Q} -divisor on X . It is known that $\kappa(D) \leq \nu(D)$ (cf. §2). D is said to be *abundant* if the equation $\kappa(D) = \nu(D)$ holds.

ABUNDANCE CONJECTURE. If X is a minimal variety, then K_X is abundant. We say that X is a *good* minimal variety if K_X is abundant.

11.1. THEOREM (KAWAMATA). *Let X be a good minimal variety. Then the canonical divisor K_X is semiample. So there is a contraction $f: X \rightarrow Z$ such that*

- (i) $\dim Z = \nu(K_X)$,
- (ii) $K_X|_{f^{-1}(z)} \equiv 0$ for all $z \in Z$.

The abundance conjecture is true when $\nu(K_X) = \dim X$, $\nu(K_X) = 0$ and when $\kappa(K_X) = \kappa(X) = \dim X - 1$. Moreover, we have

11.2. THEOREM (KAWAMATA). *The model and the abundance conjectures for dimension $\leq \dim X - k$ imply that a minimal variety X is good when $\kappa(X) \geq k$.*

So if the model conjecture is true we have only to establish the following statement in order to prove the abundance conjecture:

WEAK ABUNDANCE CONJECTURE. If X is a minimal variety and $\nu(K_X) \geq 1$, then $\kappa(X) \geq 1$.

We have only the following result in dimension three.

11.3. THEOREM (MIYAOKA). *Let X be a minimal variety and $\dim X = 3$. Then $\kappa(X) \geq 0$.*

12. Bibliographical notes. A complete account of the current state in the minimal model program and its applications can be found in [KMM].

The notion of a canonical and a terminal singularity and that of a canonical model were introduced by Reid in [R1]. Minimal models in dimension three emerge and are constructed from canonical models in [R2]. Of the algebraic 3-folds, nonsingular Fano 3-folds seem to be the only class that has been studied extensively in recent years (see [I, MM] and their references). In the singular three-dimensional case we have only one essential result about Fano 3-folds, which says that if X is a Gorenstein Fano 3-fold then a general divisor in the anticanonical linear system $|-K_X|$ is a $K3$ surface, possibly with canonical singularities [R3]. An important result on birational automorphisms of conic bundles was obtained by Sarkisov [S].

The nonvanishing and the generalized vanishing theorem originally proved in [S2, K1, V] are also given in [Kol1, K2, K3]. The contraction theorem was first proved only in dimension ≤ 3 , in the case where the nonvanishing theorem follows immediately from the Riemann-Roch theorem [K4, B, K5, S3]. For any dimension the contraction theorem was proved in [S2], where the author improved the technique developed by Kawamata, Benveniste, and Reid [R3].

The first proof of the cone theorem for nonsingular varieties [M] has used the ingenious method of modulo p reduction. Then in [K5, R3, S4] for dimension ≤ 3 the cone theorem was derived from a combination of the contraction theorem and the rationality theorem, which was proved by the same technique as that used in the proof of the contraction theorem. Improving this technique Kawamata and Kollár [K3, Kol2] have proved the theorem in the general case.

The flip conjecture was first stated and proved for toric morphisms in [R4]. General properties of flips can be found in [S2]. Sufficient conditions of semistable type (cf. Theorem 9.1) were first introduced by Tsunoda [T]. Other approaches were developed by Mori, Kawamata [K6], and by the author [S5]. Kawamata's and some other conditions are described in [K6].

Termination in dimension three was proved in [S2] and in dimension four in [KMM].

The abundance conjecture and related results are discussed in [K2]. For the last proof of the Miyoka theorem see [Mi].

Note added in proof. Now the flip and the model conjectures hold in dimension 3; see S. Mori, *Flip theorem and the existence of minimal models for 3-folds*, Preprint.

REFERENCES

- [B] X. Benveniste, *Sur l'anneau canonique de certaines variétés de dimension 3*, Invent. Math. **73** (1983), 157–164.
- [I] V. A. Iskovskih, *Algebraic threefolds with special regard to the problem of rationality*, Proc. Internat. Congr. Math. (Warsaw, 1982), PWN, Warsaw, 1983, pp. 733–746.
- [K1] Y. Kawamata, *A generalization of Kodaira-Ramanujam's vanishing theorem*, Math. Ann. **261** (1982), 43–46.
- [K2] —, *Pluricanonical systems on minimal algebraic varieties*, Invent. Math. **79** (1985), 567–588.
- [K3] —, *The cone of curves of algebraic varieties*, Ann. of Math. (2) **119** (1984), 603–633.
- [K4] —, *On the finiteness of generators of a pluricanonical ring for a 3-fold of general type*, Amer. J. Math. **106** (1984), 1503–1512.
- [K5] —, *Elementary contractions of algebraic 3-folds*, Ann. of Math. (2) **119** (1984), 95–110.
- [K6] —, *Crepant blowing-ups of 3-dimensional canonical singularities and their application to degenerations of surfaces*, Preprint, 1986.
- [KMM] Y. Kawamata, K. Matsuda, and K. Matsuki, *Introduction to the minimal model problem*, Adv. Stud. Pure Math. (to appear).
- [Kl] S. Kleiman, *Toward a numerical theory of ampleness*, Ann. of Math. **84** (1966), 293–344.
- [Kol1] J. Kollár, *Higher direct images of dualizing sheaves*, Preprint, Harvard Univ., 1984.
- [Kol2] —, *The cone theorem*, Note to a paper: "The cone of curves of algebraic varieties" by Y. Kawamata [K3], Ann. of Math. (2) **120** (1984), 1–5.
- [M1] Y. Miyaoka, *The Kodaira dimension of minimal threefolds*, Preprint, 1986.
- [M] S. Mori, *Threefolds whose canonical bundles are not numerically effective*, Ann. of Math. (2) **116** (1982), 133–176.
- [MM] S. Mori and S. Mukai, *Classification of Fano 3-folds with $B_2 \geq 2$* , Manuscripta Math. **36** (1981), 147–162.
- [R1] M. Reid, *Canonical 3-folds*, Géométrie Algébrique Angers 1979, A. Beauville, editor, Sijthoff & Noordhoff, Alphen aan den Rijn, 1980, pp. 273–310.
- [R2] —, *Minimal models of canonical 3-folds*, Algebraic Varieties and Analytic Varieties, S. Itaka, editor, Adv. Stud. Pure Math., Vol. 1, Kinokuniya, Tokyo, and North-Holland, Amsterdam, 1983, pp. 131–180.
- [R3] —, *Projective morphisms according to Kawamata*, Preprint, Warwick Univ., 1983.
- [R4] —, *Decomposition of toric morphisms*, Arithmetic and Geometry II, M. Artin and J. Tate, editors, Progr. Math., vol. 36, Birkhäuser, 1983, pp. 395–418.
- [S] V. G. Sarkisov, *On conic bundle structures*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), 371–408=Math. USSR Izv. **20** (1983), 355–390.
- [S1] V. V. Shokurov *A new proof of the Kulikov theorem* (to appear).
- [S2] —, *The nonvanishing theorem*, Izv. Akad. Nauk SSSR Ser. Mat. **49** (1985), 635–651=Math. USSR Izv. **26** (1986), 591–604.
- [S3] —, *Extremal contractions of algebraic 3-folds*, Birational Geometry of Algebraic Varieties (Proc. Conf.), Yaroslav. Gos. Ped. Inst., Yaroslavl', 1983, pp. 74–90. (Russian)
- [S4] —, *On the closed cone of curves of algebraic 3-folds*, Izv. Akad. Nauk SSSR Ser. Mat. **48** (1984), 203–208=Math. USSR Izv. **24** (1985), 193–198.
- [S5] —, *Relative extremal modifications*, Preprint, 1985. (Russian)
- [T] S. Tsunoda, *Degenerations of surfaces*, Adv. Stud. Pure Math. (to appear).
- [V] E. Viehweg, *Vanishing theorems*, J. Reine. Angew. Math. **335** (1982), 1–8.

Vanishing Theorems and Positivity in Algebraic Fibre Spaces

ECKART VIEHWEG

This note tries to explain some methods used in and initiated by the birational classification theory of complex projective manifolds. We want to draw attention to improvements of vanishing theorems and to properties of direct images of certain sheaves which should be of some interest outside of the birational geometry as well. First, however, we have to recall briefly the open problems and the state of the art in classification theory. (See the survey articles [1–7] for a more complete description.)

1. Minimal models. S. Mori's theory of the cone of curves and its extremal rays, as well as his contraction theorem [22], gave some hope that any complex projective manifold X might have a (singular) minimal model.

As indicated by M. Reid's structure theorems [5] on the canonical model of threefolds, one has to allow at least "canonical singularities" (Conditions 1 and 2):

DEFINITION 1. A normal projective variety V is called a minimal variety if there exists some $r > 0$ such that:

- (1) The reflexive hull $\omega_V^{[r]}$ of the r th power of the canonical sheaf is invertible.
- (2) For all (or one) desingularization $\delta: X \rightarrow V$, one has $\omega_X^r \supset \delta^* \omega_V^{[r]}$.
- (3) $\omega_V^{[r]}$ is numerically effective; i.e.,

$$\deg(\omega_V^{[r]}|_C) \geq 0$$

for all curves C in V .

(The minimal number r satisfying (1) is called the *index* of V .)

THE MINIMAL MODEL CONJECTURE. *Any non-uniruled complex projective manifold X is birationally equivalent to a minimal variety V (called the minimal model).*

Added in proof. Shigefumi Mori recently gave a proof of the minimal model conjecture for threefolds.

Even though S. Mori does not like the word “conjecture” in this connection, the quite encouraging results of Y. Kawamata and himself (for example, [22, 18]) seem to justify this expression. In fact, most results mentioned in the sequel assume that this conjecture holds.

2. Semi-ampleness.

DEFINITION 2. Let V be a normal projective variety and \mathcal{L} an invertible sheaf.

- (a) \mathcal{L} is called semi-ample if for some $\mu > 0$, \mathcal{L}^μ is generated by $H^0(V, \mathcal{L}^\mu)$.
- (b) The Iitaka dimension of \mathcal{L} is

$$\kappa(V, \mathcal{L}) := \begin{cases} -\infty & \text{if } R(V, \mathcal{L}) = \mathbf{C}, \\ \operatorname{tr} \deg_{\mathbf{C}} R(V, \mathcal{L}) - 1 & \text{otherwise,} \end{cases}$$

where $R(V, \mathcal{L}) := \bigoplus_{i \geq 0} H^0(V, \mathcal{L}^i)$.

DEFINITION 3. If V is nonsingular (or a minimal model of index r), then the Kodaira dimension of V is

$$\kappa(V) := \kappa(V, \omega_V) \quad (\text{or } \kappa(V) := \kappa(V, \omega_V^{[r]}).$$

SEMI-AMPLENESS CONJECTURE. *If V is a minimal variety of index r , then $\omega_V^{[r]}$ is semi-ample.*

Using the vanishing theorems of §4, this conjecture has been verified by X. Benveniste, Y. Kawamata, and V. Shokurov if $\kappa(V) = \dim V$ and $\kappa(V) = \dim V - 1$ (see [16, 18] for details). Of course, the conjecture implies that $\kappa(V) \geq 0$ for all minimal varieties V . By studying deformations of subvarieties along foliations, Y. Miyaoka [21] recently verified this implication directly in the 3-dimensional case.

3. The Iitaka-Ueno program ([13, 24]). The starting point of the recent development in classification theory was S. Iitaka’s fibration theorem, saying roughly that each manifold with $\kappa(X) \geq 0$ carries a fibration by subvarieties of dimension $\dim X - \kappa(X)$ and of Kodaira dimension zero. In a similar way, one would like to use the Albanese map to reduce the study of manifolds X with $\kappa(X) \leq 0$ and $q(X) = \dim H^0(X, \Omega_X^1) > 0$ to families of lower-dimensional manifolds. In order to do so, one needs an affirmative answer to the

GENERALIZED IITAKA CONJECTURE. *For any fibre space $f: X \rightarrow Y$, with $\kappa(Y) \geq 0$, one has $\kappa(X) \geq \operatorname{Max}\{\operatorname{Var}(f), \kappa(Y)\} + \kappa(X_y)$.*

We use

DEFINITION 4. (a) A surjective morphism $f: X \rightarrow Y$ between projective manifolds is called a *fibre space* if the general fibre X_y is connected.

(b) $\operatorname{Var}(f)$ is the minimal transcendental degree over \mathbf{C} of a field of definition for the birational equivalence class of X_y .

As explained in §7, this conjecture is now solved by J. Kollár, if $\kappa(X_y) = \dim X_y$ [19], and by Kawamata if X_y has a minimal model satisfying the Semi-ampleness Conjecture [15]. However, it can be applied successfully to the Albanese map only, if the structure of X_y is sufficiently well known. (See [15, §8] and [25, §9] for typical applications.)

The Iitaka Conjecture, the Semi-ampleness Conjecture, and even some of the methods used to attack the Minimal Model Conjecture are dealing with the general problem: “How to get lots of sections of powers of canonical sheaves.” It is not too surprising that the same type of methods appears in most results obtained in one of the three directions.

4. Vanishing theorems. One method of studying sections (or cohomology) of an invertible sheaf \mathcal{M} on a projective manifold X consists of restricting it to an effective divisor B and hoping that the surjectivity of the restriction map allows induction on the dimension. Hence, if we write $\mathcal{M} = \mathcal{L} \otimes \omega_X(B) := \mathcal{L} \otimes \omega_X \otimes \mathcal{O}_X(B)$, we ask for criteria implying that the adjunction map

$$\phi_{q,B}: H^q(X, \mathcal{L} \otimes \omega_X(B)) \rightarrow H^q(B, \mathcal{L} \otimes \omega_B)$$

is surjective. As a generalization of Kodaira’s vanishing theorem, Kawamata and I obtained (see [18, 1-2] or [8, 2.13])

THEOREM 5. *Let $C = \sum r_i C_i$ be an effective normal crossing divisor such that for some $N > \max\{r_i\}$, $\mathcal{L}^N(-C)$ is numerically effective. Then*

$$H^q(\mathcal{L} \otimes \omega_X) = 0$$

(and hence $\phi_{q-1,B}$ is surjective for all B) for $q > \dim X - \kappa(X, \mathcal{L})$.

Usually Theorem 5 is called “the vanishing theorem for integral parts of \mathbf{Q} -divisors” to show that the “positivity assumption” is just made for $\mathcal{L}(-\frac{1}{N} \cdot C)$. This tiny improvement turns out to be quite important for applications. A more direct approach to study the adjunction map was given by Kollár [20]:

THEOREM 6. *If \mathcal{L}^N is generated by its global sections and $\mathcal{O}_X(B) \subset \mathcal{L}^N$ for some $N \gg 0$, then $\phi_{q,B}$ is surjective for all q .*

In fact, both Theorems 5 and 6 can be obtained using the degeneration of the Hodge-Deligne spectral sequence. For example, if B is smooth and $\mathcal{O}_X(B) = \mathcal{L}^N$, the proof of (6) given in [8] reads:

Let $\pi: X' \rightarrow X$ be the cyclic cover obtained by taking the N th root out of a section of \mathcal{L}^N defining B . Then \mathcal{L}^{-1} is a direct summand of $\pi_* \mathcal{O}_{X'}$, and the differential $d: \mathcal{O}_{X'} \rightarrow \Omega_{X'}^1 \langle B \rangle$ induces an integrable connection $\nabla: \mathcal{L}^{-1} \rightarrow \mathcal{L}^{-1} \otimes \Omega_{X'}^1 \langle B \rangle$. For the Poincaré residue $r: \Omega_X^1 \langle B \rangle \rightarrow \mathcal{O}_B$, $(\text{id} \otimes r) \circ \nabla$ is nothing but the restriction map $\mathcal{L}^{-1} \rightarrow \mathcal{L}^{-1}|_B$.

By Deligne’s mixed Hodge theory, $H^q(d) = 0$ and hence

$$0 = H^q((\text{id} \otimes r) \circ \nabla): H^q(\mathcal{L}^{-1}) \rightarrow H^q(\mathcal{L}^{-1}|_B).$$

A similar proof gives a slight generalization of Theorem 6:

THEOREM 7. *If $\mathcal{L}^N(-C) = \mathcal{O}_X$ for C and N as in Theorem 5, then $\phi_{q,B}$ is surjective for all divisors B with $B_{\text{red}} \leq C_{\text{red}}$.*

Besides the applications of Theorems 5 and 6 to the Semi-ampleness Conjecture mentioned already, and the simple proof of the “Positivity Theorem” [20], described in §5, the two statements were successfully applied

- to the deformation behavior of plurigenera [23],
- in deformation theory of singularities [14 and 12],
- for Nullstellen Lemmata for homogeneous polynomials [10],
- and even, together with methods described in §5, to reprove the theorem of Roth on diophantine approximations of algebraic numbers [11].

5. Weak positivity. Let $f: X \rightarrow Y$ be a fibre space. In order to construct sections of powers of $\mathcal{L} \otimes \omega_X$, one can try to study positivity properties of $f_*(\mathcal{L} \otimes \omega_{X/Y})$, where $\omega_{X/Y} := \omega_X \otimes f^*\omega_Y^{-1}$. In order to have a notation of positivity compatible with tensor products, symmetric products, finite covers or blowing ups, one defines

DEFINITION 7. Let \mathcal{F} be a sheaf on Y .

(a) \mathcal{F} is called *generically generated* if $H^0(Y, \mathcal{F}) \otimes_{\mathbb{C}} \mathcal{O}_Y \rightarrow \mathcal{F}$ is surjective at the general point of Y .

(b) \mathcal{F} is called *weakly positive* if for every ample invertible sheaf \mathcal{H} on Y and every $a > 0$ there exists some $b > 0$ such that $\hat{S}^{a+b}(\mathcal{F}) \otimes \mathcal{H}^b$ is generically generated (where \hat{S} denotes the reflexive hull of the symmetric product).

(c) \mathcal{F} is called *big* if for all ample invertible sheaves \mathcal{H} there is some $c > 0$ such that $\mathcal{H}^{-1} \otimes S^c(\mathcal{F})$ is weakly positive.

For example, if \mathcal{F} is invertible, \mathcal{F} is weakly positive if and only if the corresponding divisor is in the closure of the cone of effective divisors in $\text{Pic}(Y)$; and \mathcal{F} is big if and only if $\kappa(Y, \mathcal{F}) = \dim Y$. A slight modification of the positivity theorem of Fujita-Kawamata [17, 25] says:

THEOREM 8. *If $f: X \rightarrow Y$ is a fibre space, then $f_*\omega_{X/Y}$ is weakly positive.*

Kollár [20] realized that the quite difficult study on variations of Hodge structures used by Kawamata to prove Theorem 8 can be replaced by his vanishing theorem (6), and it is even sufficient to consider a smooth divisor B with $\mathcal{O}_X(B) \subset \mathcal{L}^N$. Moreover, his approach gives an effective bound of the number b in the definition of weakly positive -independent of a :

THEOREM 8'. *Let \mathcal{H} be a very ample invertible sheaf on Y . Then for $m > \dim Y$ and all $s > 0$ the sheaf $\mathcal{H}^m \otimes \omega_Y \otimes \hat{S}^s(f_*\omega_{X/Y})$ is generically generated.*

SKETCH OF THE PROOF. If $f^{(s)}: X^{(s)} \rightarrow Y$ is a desingularization of the s -fold product $X \times_Y X \times_Y \cdots \times_Y X$, there is a map $f_*^{(s)}\omega_{X^{(s)}/Y} \rightarrow (f_*\omega_{X/Y})^{\otimes s}$ which is generically an isomorphism. Hence it is enough to consider $s = 1$. If H is a general divisor of \mathcal{H} , $\mathcal{L} = f^*\mathcal{H}$ and $B = f^{-1}(H)$, Theorem 6 gives a

surjection

$$\begin{aligned} H^0(\mathcal{H}^m \otimes \omega_Y \otimes f_* \omega_{X/Y}) &= H^0(\mathcal{L}^{m-1} \otimes \omega_X(B)) \rightarrow H^0(\mathcal{L}^{m-1} \otimes \omega_B) \\ &= H^0(\mathcal{H}^{m-1} \otimes \omega_H \otimes f_* \omega_{B/H}), \end{aligned}$$

and by induction on $\dim Y$ we are done.

Applying Theorem 8 to a cyclic cover of X and playing around with the properties of weakly positive sheaves, one obtains [25]

THEOREM 9. (a) *Let $C = \sum r_i C_i$ be an effective normal crossing divisor, \mathcal{L} an invertible sheaf on X , and $N > \max\{r_i\}$ such that $\mathcal{L}^N(-C)$ is generated by its global sections. Then $f_*(\mathcal{L} \otimes \omega_{X/Y})$ is weakly positive.*

(b) *For all $\nu > 0$ the sheaf $f_* \omega_{X/Y}^\nu$ is weakly positive.*

If Y is of general type, Theorem 9(b) implies the Iitaka Conjecture ($\kappa(X) = \kappa(Y) + \kappa(X_y)$).

The effective bound of Kollár carries over, and if $\eta(Z)$ denotes the smallest number for which the η th canonical map of a manifold Z is generically finite, one obtains [9]

$$\eta(X) \leq \eta(X_y) \cdot \left\lceil \frac{(m+2) \cdot \eta(Y) + 2}{\eta(X_y)} \right\rceil$$

where $\lceil \cdot \rceil$ denotes the “round up” of a real number.

6. Big sheaves. The sheaves $f_* \omega_{X/Y}^\nu$ satisfy some kind of generic stability condition if $\nu > 1$ (see [25 and 26]). For example, if Y is a curve and if $f_* \omega_{X/Y}^\nu$ contains an ample invertible sheaf, then $f_* \omega_{X/Y}^\nu$ is ample. Obviously a similar statement is wrong for $\nu = 1$.

If the degenerated fibres of f are not too bad one obtains

THEOREM 10. *If $f: X \rightarrow Y$ is a fibre space and if the fibres of f are semistable outside a subvariety of Y of codimension two, then for all $\eta > 1$ one has*

$$\kappa(\det(f_* \omega_{X/Y}^\eta)) = \dim Y \quad \text{if and only if } f_* \omega_{X/Y}^\eta \text{ is big.}$$

Of course, in order to have the bigness of $f_* \omega_{X/Y}^\eta$, the fibre space f should have the maximal number of moduli. Using the notations introduced in Definition 4(b) one hopes that:

BIGNESS PROBLEM. *Does $\kappa(X_y) \geq 0$ and $\text{Var}(f) = \dim Y$ imply that $f_* \omega_{X/Y}^\eta$ is big for some $\eta \gg 0$?*

Similar to the Minimal Model Conjecture and the Semi-ampleness Conjecture one might even be tempted to ask, in addition, whether there is a birational model for the fibre space such that $f_* \omega_{X/Y}^\eta$ is generated by its global sections. As explained in [25] and [26], an affirmative answer to the Bigness Problem gives a solution of the Generalized Iitaka Conjecture (§3).

Theorem 10 can be used together with a quite technical covering construction to reduce the Bigness Problem for $f: X \rightarrow Y$ to that for a fibre space

$f': X' \rightarrow Y'$ whose general fibre X'_y is obtained as a cyclic cover of X_y , ramified along a multicanonical divisor. (See [26 and 19] for an extended version.)

7. Moduli. If $f: X \rightarrow Y$ is a family of curves or surfaces of general type, it follows from D. Mumford's and D. Gieseker's construction of a projective moduli-variety that for $k, s \gg 0$ and $r(k) = \text{rank}(f_*\omega_{X/Y}^k)$ one has

$$\kappa(Y, \det(f_*\omega_{X/Y}^{k \cdot s})^{\cdot r(k)} \otimes \det(f_*\omega_{X/Y}^k)^{-s \cdot r(s \cdot k)}) \geq \text{Var}(f).$$

Together with Theorem 9(b) this implies an affirmative answer to the Bigness Problem [25].

Instead of global moduli one can as well use local Torelli theorems for the general fibre X_y . By the remark at the end of the last section one has to have Torelli-type theorems only for certain coverings of X_y . Those are quite easily verified if $\kappa(X_y) = \dim X_y$ and if ω_{X_y} is semi-ample [27]. Kawamata considered this approach under much weaker assumptions and obtained, in [15],

THEOREM 11. *The bigness problem has an affirmative answer if X_y has a minimal model satisfying the semi-ampleness conjecture.*

Kollár used in [19] properties of variations of Hodge structures to get hold of the kernel of the multiplication map $\gamma_m: S^m(f_*\omega_{X/Y}) \rightarrow f_*\omega_{X/Y}^m$. If $\text{Var}(f) = \dim Y$ and if the canonical map of X_y is birational, he obtained that the image of γ_m is "positive." Together with the usual covering construction this implies

THEOREM 12. *The bigness problem has an affirmative answer if $\kappa(X_y) = \dim X_y$.*

REFERENCES

Survey articles (containing more references)

1. H. Esnault, *Classification des variétés de dimension 3 et plus*, Bourbaki Seminar (February 1981), Lecture Notes in Math., Vol. 901, Springer-Verlag, Berlin-New York, 1981, pp. 111–131.
2. S. Iitaka, *Birational geometry of algebraic varieties*, Proc. Internat. Congr. Math. (Warsaw, 1982), PWN, Warsaw, and North-Holland, Amsterdam-New York-Oxford, 1984, pp. 727–732.
3. S. Mori, *Cone of curves, and Fano 3-folds*, Proc. Internat. Congr. Math. (Warsaw, 1982), PWN, Warsaw, and North-Holland, Amsterdam-New York-Oxford, 1984, pp. 747–752.
4. —, *Classification of higher dimensional varieties* (Proc. AMS Summer Institute, Bowdoin College, 1985) (to appear).
5. M. Reid, *Young person's guide to canonical singularities* (Proc. AMS Summer Institute, Bowdoin College, 1985) (to appear).
6. K. Ueno, *Classification of algebraic manifolds*, Proc. Internat. Congr. Math. (Helsinki, 1978), Acad. Sci. Fennica, Helsinki, 1980, pp. 549–556.
7. E. Viehweg, *Zur Klassifikationstheorie drei (und höher) dimensionaler projektiver Mannigfaltigkeiten*, Jahresber. Deutsch. Math.-Verein **86** (1984), 135–150.

Articles and lecture notes

8. H. Esnault and E. Viehweg, *Logarithmic De Rham complexes and vanishing theorems*, Invent. Math. **86** (1986), 161–194.
9. —, *Revêtements cycliques*. II (Proceedings, La Rábida, 1984) (to appear).

10. —, *Sur une minoration du degré d'hypersurfaces s'annulant en certains points*, Math. Ann. **263** (1983), 75–86.
11. —, *Dyson's Lemma for polynomials in several variables (and the Theorem of Roth)*, Invent. Math. **78** (1984), 445–490.
12. —, *Two dimensional quotient singularities deform to quotient singularities*, Math. Ann. **271** (1985), 439–449.
13. S. Iitaka, *On D -dimension of algebraic varieties*, J. Math. Soc. Japan **23** (1971), 356–373.
14. S. Ishii, *Small deformations of normal singularities*, Math. Ann. **275** (1986), 139–148.
15. Y. Kawamata, *Minimal models and the Kodaira dimension of algebraic fibre spaces*, J. Reine Angew. Math. **363** (1985), 1–46.
16. —, *Pluricanonical systems on minimal algebraic varieties*, Invent. Math. **79** (1985), 567–588.
17. —, *Characterization of abelian varieties*, Compositio Math. **43** (1981), 253–276.
18. Y. Kawamata, K. Matsuda, and K. Matsuki, *Introduction to the minimal model problem*, Preprint, Tokyo Univ.
19. J. Kollár, *Subadditivity of the Kodaira dimension: fibers of general type*, Preprint, Harvard Univ.
20. —, *Higher direct images of dualizing sheaves. I*, Ann. of Math. **123** (1986), 11–42.
21. Y. Miyaoka, *On the Kodaira dimension of minimal threefolds*, Preprint, Tokyo Metropolitan Univ.
22. S. Mori, *Threefolds whose canonical bundles are not numerically effective*, Ann. of Math. **116** (1982), 133–176.
23. N. Nakayama, *Invariance of the plurigeners of algebraic varieties under minimal model conjectures*, Topology **25** (1986), 237–251.
24. K. Ueno, *Classification theory of algebraic varieties and compact complex spaces*, Lecture Notes in Math., Vol. 439, Springer-Verlag, Berlin-Heidelberg-New York, 1975.
25. E. Viehweg, *Weak positivity and the additivity of the Kodaira dimension for certain fibre spaces*, Algebraic Varieties and Analytic Varieties (Tokyo, 1981), Adv. Stud. Pure Math., no. 1, North-Holland, Amsterdam, 1983, pp. 329–353.
26. —, *Weak positivity and the additivity of the Kodaira dimension II: The local Torelli map*, Classification of Algebraic and Analytic Manifolds (Katata, 1982), Progress in Math., Vol. 39, Birkhäuser, Boston-Basel-Stuttgart, 1983, pp. 567–589.

UNIVERSITÄT GHS ESSEN, FB 6 MATHEMATIK, D 4300 ESSEN 1, FEDERAL REPUBLIC OF GERMANY

L-Series and the Green's Functions of Modular Curves

DON ZAGIER

The problem we want to discuss in this paper is the following:

Construct on the modular curve $X_0(N)$ explicit divisors of degree 0 defined over \mathbf{Q} , and relate their heights to the derivatives at $s = 1$ of L -series of cusp forms of weight 2 and level N . (1)

We begin by recalling the definitions of the various terms occurring here and the motivation for the question.

Let X be a curve defined over \mathbf{Q} . By a *divisor of degree 0* on X we mean a finite formal linear combination $\mathfrak{x} = \sum_i n_i(x_i)$ ($x_i \in X(\overline{\mathbf{Q}})$, $n_i \in \mathbf{Z}$) with $\sum_i n_i = 0$. It is *defined over \mathbf{Q}* if $\mathfrak{x}^\sigma = \mathfrak{x}$ for all $\sigma \in \text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. The quotient of the abelian group of such divisors by the subgroup of principal divisors

$$(f) = \sum_{\substack{x \text{ a zero or} \\ \text{pole of } f}} \text{ord}_x(f) \cdot (x) \quad (f: X \rightarrow \mathbf{P}^1 \text{ a rational function defined over } \mathbf{Q})$$

is (if X has a \mathbf{Q} -rational point) the set $J_X(\mathbf{Q})$ of rational points on a certain abelian variety J_X , the Jacobian of X , and is a finitely generated group by the Mordell-Weil theorem. The Néron-Tate theory associates to each \mathfrak{x} a real number $h(\mathfrak{x}) \geq 0$, called its *height* (or canonical height), which depends only on the class of \mathfrak{x} in $J_X(\mathbf{Q})$ and which defines a quadratic form on $J_X(\mathbf{Q})$, i.e., $h(\mathfrak{x}) = \langle \mathfrak{x}, \mathfrak{x} \rangle$ for a certain symmetric bilinear form $\langle \cdot, \cdot \rangle$ on $J_X(\mathbf{Q})$ (the *height pairing*). Moreover, h is positive definite on the free abelian group $J_X(\mathbf{Q})/\text{torsion}$; i.e., $h(\mathfrak{x})$ vanishes only if \mathfrak{x} is of finite order in $J_X(\mathbf{Q})$. We will explain later how the height is defined.

For $N \in \mathbf{N}$, the *modular curve* $X_0(N)$ is a curve defined over \mathbf{Q} whose \mathbf{C} -rational points are given by

$$X_0(N)(\mathbf{C}) = \mathfrak{H}/\Gamma_0(N) \cup (\text{finite set of "cusps"});$$

here $\mathfrak{H} = \{z = x + iy | y > 0\}$ is the upper half-plane and $\Gamma_0(N)$ the group of matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{SL}_2(\mathbf{Z})$ with $c \equiv 0 \pmod{N}$, acting in the usual way $z \mapsto \frac{az+b}{cz+d}$ on \mathfrak{H} . The points of $X_0(N)$ over a subfield $k \subset \mathbf{C}$ parametrize pairs (E, C) consisting of an elliptic curve E and a cyclic subgroup $C \subset E(\mathbf{C})$ of order

N , both defined over k . A *cusp form of weight 2 and level N* is a holomorphic function $f: \mathfrak{H} \rightarrow \mathbf{C}$ such that the differential form $f(z) dz$ on \mathfrak{H} is $\Gamma_0(N)$ -invariant and satisfying $f(z) = O(y^{-1})$ in \mathfrak{H} . The set of such forms is a finite-dimensional vector space which we will denote by $S_2(N)$. Each $f \in S_2(N)$ has a convergent Fourier development

$$f(z) = \sum_{n=1}^{\infty} a_f(n) e^{2\pi i n z} \quad (z \in \mathfrak{H}).$$

The L -series of f is the associated Dirichlet series

$$L(f, s) = \sum_{n=1}^{\infty} a_f(n) n^{-s} \quad (s \in \mathbf{C});$$

it converges for $\operatorname{Re}(s) > \frac{3}{2}$ and has a holomorphic continuation to all s . The space $S_2(N)$ has a basis of special modular forms f (*Hecke forms*) whose L -series have Euler products (in particular, $a_f(1) = 1$) and satisfy a functional equation

$$\left(\frac{\sqrt{N}}{2\pi}\right)^s \Gamma(s) L(f, s) = w_f \left(\frac{\sqrt{N}}{2\pi}\right)^{2-s} \Gamma(2-s) L(f, 2-s)$$

with $w_f = \pm 1$. We will be particularly interested in the value of the derivative $L'(f, 1)$ for those Hecke forms with $w_f = -1$ (and hence $L(f, 1) = 0$).

The motivation for formulating the problem (1) comes from the conjecture of Birch and Swinnerton-Dyer. Given a Hecke form f whose Fourier coefficients $a_f(n)$ are all rational integers, there exists an elliptic curve E defined over \mathbf{Q} and related to f by

$$a_f(p) + \#E(\mathbf{Z}/p\mathbf{Z}) = p + 1 \quad \text{for all } p \nmid N$$

(Eichler, Shimura); this curve is unique up to isogeny (Faltings). Conversely, it is conjectured that every elliptic curve over \mathbf{Q} arises in this way (Taniyama, Weil). The Birch–Swinnerton-Dyer conjecture predicts that if $L(f, 1)$ vanishes (in particular, if $w_f = -1$), then $E(\mathbf{Q})$ has infinite order and

$$L'(f, 1) = h(P) \cdot \omega \quad \text{for some } P \in E(\mathbf{Q}) \otimes \mathbf{Q}, \quad (2)$$

where ω is an explicitly specified positive real number (a certain period of E) and $h: E(\mathbf{Q}) \rightarrow \mathbf{R}$ is the height function on E , identified with its own Jacobian via $x \mapsto (x) - (0)$. Now the point is that there is a nonconstant map $\phi: X_0(N) \rightarrow E$ defined over \mathbf{Q} . (Over \mathbf{C} , ϕ is given as follows; by the $\Gamma_0(N)$ -invariance of $f(z) dz$, the function $F(z) = \sum_{n=1}^{\infty} n^{-1} a_f(n) e^{2\pi i n z}$ satisfies $F(\gamma z) = F(z) + c_\gamma$ for all $\gamma \in \Gamma_0(N)$; the c_γ all lie in a certain 2-dimensional lattice $\Lambda \subset \mathbf{C}$, with $E(\mathbf{C}) = \mathbf{C}/\Lambda$, so F induces a map $\phi: \mathfrak{H}/\Gamma_0(N) \rightarrow E(\mathbf{C})$, and this map extends smoothly over the cusps.) Hence to any \mathbf{Q} -rational divisor $\mathfrak{x} = \sum n_i(x_i)$ on $X_0(N)$, we can associate a \mathbf{Q} -rational divisor $\sum n_i(\phi(x_i))$ and hence—since E is a group—a \mathbf{Q} -rational point $P = \sum n_i \phi(x_i)$ on E , the heights of P and \mathfrak{x} being related in a simple way. In this way a solution of (1) can be used to prove (2). We now proceed to describe one such solution.

We must first construct a \mathbf{Q} -rational divisor on $X_0(N)$. The construction is based on the theory of complex multiplication and is due to Heegner and Birch (cf. [1, 4]). We need an auxiliary piece of data. This will be an imaginary quadratic field K whose discriminant D is assumed to be prime to $2N$ and congruent to a square modulo $4N$ (equivalently, every prime divisor of N should split in K). Then there are infinitely many $\tau \in \mathfrak{H}$ satisfying a quadratic equation

$$a\tau^2 + b\tau + c = 0, \quad a, b, c \in \mathbf{Z}, \quad b^2 - 4ac = D, \quad N|a,$$

but only finitely many modulo the action of $\Gamma_0(N)$, say $\tau_{D,1}, \dots, \tau_{D,h} \in \mathfrak{H}/\Gamma_0(N) \subset X_0(N)(\mathbf{C})$ (h is a certain class number). The theory of complex multiplication tells us that the $\tau_{D,i}$ are algebraic in $X_0(N)(\mathbf{C})$ and are permuted by $\text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, so the divisor of degree zero $\mathfrak{x}_D = \sum_{i=1}^h (\tau_{D,i}) - h \cdot (\infty)$ is defined over \mathbf{Q} (here " ∞ " denotes the cusp at infinity on $X_0(N)$, which is defined over \mathbf{Q}). If D is -3 or -4 , we must divide this by 3 or 2 to get consistent formulas later, because of the presence of extra units in K ; the class of \mathfrak{x}_D is then in $J_{X_0(N)}(\mathbf{Q}) \otimes \mathbf{Q}$. We call \mathfrak{x}_D the *Heegner divisor* attached to K . If f is a Hecke form with integral Fourier coefficients associated to an elliptic curve E/\mathbf{Q} and a map $\phi: X_0(N) \rightarrow E$ as above, then we get a point

$$P_D = \sum_{i=1}^h \phi(\tau_{D,i}) \in E(\mathbf{Q}),$$

the *Heegner point* attached to K . A solution to (1) is then provided by the following theorem.

THEOREM 1 [7, 9]. *Let D, N be as above, $h(\cdot)$ the canonical height function for $X_0(N)$ over \mathbf{Q} . Assume N is prime. Then*

$$h(\mathfrak{x}_D) = \frac{|D|^{1/2}}{4\pi^2} \sum_f \frac{1}{\|f\|^2} L'(f, 1) L(f, D, 1), \quad (3)$$

where the sum runs over all Hecke forms $f \in S_2(N)$ with $w_f = -1$, $\|f\|^2$ denotes the Petersson scalar product $\iint_{H/\Gamma_0(N)} |f(z)|^2 dx dy$, and $L(f, D, 1)$ is the value at $s = 1$ of the "twisted L -series"

$$L(f, D, s) = \sum_{n=1}^{\infty} a_f(n) \left(\frac{D}{n} \right) n^{-s} \quad (\text{Re}(s) \gg 0).$$

Actually, the theorem of [9] is proved for all N , and we have assumed N prime only for convenience in stating the result. (If N is composite, there is an extra power of 2 in the formula (3) and one has to discuss new and old forms.) It is more general than Theorem 1 in two other respects. First, it gives not only the height pairing of \mathfrak{x}_D with itself, but of the individual $(\tau_{D,i}) - (\infty)$ with one another (over their field of definition, the Hilbert class field of K). Secondly, it includes the action of the Hecke operators T_n ($n \in \mathbf{N}$, $(n, N) = 1$) on $X_0(N)$; specifically, the height pairing $\langle \mathfrak{x}_D, T_n \mathfrak{x}_D \rangle$ is given by the same expression as in (3) but with an extra factor $a_f(n)$ in the f th summand. This is important since

it allows us to use the action of the Hecke algebra on $S_2(N)$ to split up the sum and get a formula in terms of heights for each term $L'(f, 1)L(f, D, 1)$ separately. In particular, if f is a Hecke form with $w_f = -1$ corresponding to an elliptic curve E/\mathbf{Q} and P_D the D th Heegner point on E as described above, then we get the formula

$$h(P_D) = \frac{|D|^{1/2}}{\text{vol}(E)} L'(f, 1)L(f, D, 1), \quad (4)$$

where $h(\cdot)$ is now the height function on $E(\mathbf{Q})$ and $\text{vol}(E)$ the volume of a fundamental parallelogram for a certain lattice Λ with $E(\mathbf{C}) = \mathbf{C}/\Lambda$. The product $L'(f, 1)L(f, D, 1)$ equals $L'(E/K, 1)$, the derivative at $s = 1$ of the L -series of E over K . Equation (4) has several consequences, most notably:

A. If $L'(f, 1) \neq 0$ (i.e., if the order of $L(f, s)$ at $s = 1$ is one), then $E(\mathbf{Q})$ has infinite order.

B. The Birch–Swinnerton-Dyer formula (2) holds up to a nonzero rational number.

C. One gets explicit examples of Hecke forms whose L -series have a zero of order ≥ 3 at $s = 1$.

The proof of A uses a theorem of Waldspurger [16], which guarantees the existence of a D with $L(f, D, 1) \neq 0$. The assertion C is of interest because, in combination with a deep result of Goldfeld [3, 12], it leads to an effective solution of Gauss's problem of showing that there are only finitely many imaginary quadratic fields having a given class number.

Formula (3) has the disadvantage that the values of $L'(f, 1)$ in which we are interested do not occur alone, but always multiplied by a twisted L -series value $L(f, D, 1)$ for some auxiliary number D . Also, we get only partial information about the positions of the Heegner divisors \mathfrak{x}_D in the Mordell-Weil group $J_{X_0(N)}(\mathbf{Q})$, namely their lengths with respect to the height pairing metric on $J_{X_0(N)}(\mathbf{Q}) \otimes \mathbf{R}$. To understand the dependence on D , we must be able to relate different discriminants, i.e., to compute the height pairing $\langle \mathfrak{x}_D, \mathfrak{x}_{D'} \rangle$ for all D, D' , not just for $D = D'$. To state the answer, we recall that Shimura [13] defined a correspondence between modular forms of weight 2 and modular forms of weight $\frac{3}{2}$ (or, more generally, weight $2k$ and weight $k + \frac{1}{2}$). If $f \in S_2(N)$ is a Hecke form, we denote its image under this correspondence by g_f and write its Fourier development as $g_f(z) = \sum_{D < 0} c_f(D) e^{2\pi i |D|z}$ ($z \in \mathfrak{H}$). We do not recall the exact definition of forms of half-integral weight or of the Shimura correspondence here; roughly, g_f satisfies

$$g_f \left(\frac{az + b}{cz + d} \right)^2 = \pm (cz + d)^3 g_f(z)^2 \quad \forall \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(4N)$$

and the relation between g_f and f is that for all n and D the ratio of $c_f(n^2 D)$ to $c_f(D)$ is given by a simple linear combination of the Fourier coefficients $a_f(d)$, $d|n$. The form g_f is unique up to a scalar multiple and cannot be normalized in any canonical way, but it can be chosen to have all its coefficients algebraic

integers in the number field generated by the Fourier coefficients of f . We can now state

THEOREM 2 [6]. *Let N be prime, D and D' discriminants of imaginary quadratic fields which are prime to N and squares modulo $4N$. Then*

$$\langle \mathfrak{x}_D, \mathfrak{x}_{D'} \rangle = \frac{3}{4\pi} \sum_f \frac{1}{\|g_f\|^2} L'(f, 1) c_f(D) c_f(D'), \quad (5)$$

where the summation is the same as in Theorem 1.

Again the result actually proved is for arbitrary N but is more complicated to state in general, not only because the numerical factor $3/4\pi$ changes and one has to worry about old and new forms, but because one has to replace the theory of modular forms of half-integral weight by the theory of Jacobi forms [2] if one wants to get a complete result. The compatibility of Theorems 1 and 2 follows from a theorem of Waldspurger [14, 15], which states

$$L(f, D, 1) = 3\pi \frac{\|f\|^2}{\|g_f\|^2} \frac{c_f(D)^2}{|D|^{1/2}}$$

(actually, Waldspurger stated only the proportionality of $c_f(D)^2/|D|^{1/2}$ and $L(f, D, 1)$ for fixed f ; the constant was determined in [11, 10]). As before, one has a generalization of (5) involving Hecke operators (replace $\langle \mathfrak{x}_D, \mathfrak{x}_{D'} \rangle$ by $\langle \mathfrak{x}_D, T_n \mathfrak{x}_{D'} \rangle$ and insert a factor $a_f(n)$ before the f th summand) and this can be used to separate the various Hecke forms and get a formula for each $L'(f, 1)$ separately. In particular, if f corresponds to an elliptic curve E/\mathbf{Q} , we get the following

COROLLARY. *Let E/\mathbf{Q} be an elliptic curve parametrized by a cusp form f of prime level and $L(E/\mathbf{Q}, s) = L(f, s)$ its L -series. If $L(E/\mathbf{Q}, s)$ has a simple zero at $s = 1$, then the space spanned by all Heegner points in $E(\mathbf{Q}) \otimes \mathbf{Q}$ is one-dimensional; more precisely, there is a nonzero point $P_0 \in E(\mathbf{Q}) \otimes \mathbf{Q}$ such that $P_D = c_f(D)P_0$ for all D , where the $c_f(D)$ are the Fourier coefficients of a form of weight $\frac{3}{2}$ corresponding to f under the Shimura correspondence. There is an analogous result for f of composite level, involving coefficients of Jacobi forms.*

That all Heegner points lie on a line when $\text{ord}_{s=1} L(E/\mathbf{Q}, s) = 1$ would follow from—and hence provides additional support for—the Birch–Swinnerton–Dyer conjecture, which predicts that the entire Mordell–Weil group has rank one in this case. If $\text{ord}_{s=1} L(E/\mathbf{Q}, s)$ is not equal to one, of course, then all Heegner points vanish (up to torsion) by Theorem 1.



The Néron–Tate height is a sum of local contributions from all places of \mathbf{Q} , and in the proof of Theorems 1 and 2 it is necessary to compute these local heights. In the second part of the talk we discuss some of the arithmetic questions which arise in this context.

We first briefly describe the local height theory. (A good reference is [5].) Let $\mathfrak{x} = \sum n_i(x_i)$ and $\mathfrak{y} = \sum m_j(y_j)$ be \mathbf{Q} -rational divisors of degree 0 on a curve X/\mathbf{Q} , and suppose for simplicity that \mathfrak{x} and \mathfrak{y} have disjoint supports (i.e., $x_i \neq y_j$ for all i, j). Then there is a decomposition

$$\langle \mathfrak{x}, \mathfrak{y} \rangle = \langle \mathfrak{x}, \mathfrak{y} \rangle_\infty + \sum_p \langle \mathfrak{x}, \mathfrak{y} \rangle_p,$$

where the summation runs over all prime numbers p but has only finitely many nonzero terms. Each $\langle \mathfrak{x}, \mathfrak{y} \rangle_v$ (v a place of \mathbf{Q}) is a real number depending continuously on the x_i and y_j in the v -adic topology; i.e., we can compute $\langle \mathfrak{x}, \mathfrak{y} \rangle_\infty$ to any degree of accuracy if we know sufficiently accurately the position of the x_i and y_j on the Riemann surface $X(\mathbf{C})$, and $\langle \mathfrak{x}, \mathfrak{y} \rangle_p$ if we know the coordinates of the x_i and y_j modulo a sufficiently large power of p . The local height pairing is bilinear and symmetric and satisfies the identity $\langle \mathfrak{x}, \mathfrak{y} \rangle_v = \sum_i n_i \log |f(x_i)|_v$ if $\mathfrak{y} = (f)$ is a principal divisor (notice that this implies $\langle \mathfrak{x}, (f) \rangle = 0$ by the product formula for valuations, which is why the global height is well defined on $J_X(\mathbf{Q})$). These properties characterize $\langle \cdot, \cdot \rangle_v$ uniquely, since the difference of two such symbols would be a continuous homomorphism of the compact group $J_X(\mathbf{Q}_v)^2$ to \mathbf{R} . Moreover, one can find an explicit solution to these axioms by using intersection theory at finite places and potential theory at infinity. Specifically, $\langle \mathfrak{x}, \mathfrak{y} \rangle_p = -n_p(\mathfrak{x}, \mathfrak{y}) \log p$ for p finite, where $n_p(\mathfrak{x}, \mathfrak{y})$ is a rational number which is integral and nonnegative if X has good reduction at p and is then 0 unless some x_i and y_j reduce to the same point modulo p . At infinity we have

$$\langle \mathfrak{x}, \mathfrak{y} \rangle_\infty = \sum_{i,j} n_i m_j G(x_i, y_j),$$

where G is a *Green's function* on X , characterized by the property

$$G(\cdot, y) \text{ is continuous and harmonic on } X \text{ except for logarithmic singularities of residue } +1 \text{ and } -1 \text{ at } y \text{ and } y_0, \text{ respectively.} \quad (6)$$

Here y_0 is a chosen basepoint on $X(\mathbf{C})$, and by "logarithmic singularity of residue c at a point P " we mean a function which looks like $c \log |z|^2 + O(1)$ near P , where z is a local uniformizing parameter with $z(P) = 0$. (Notice that the function $G(\cdot, y)$ is well defined up to a constant for fixed choice of y_0 , since the difference of any two G 's would be harmonic and finite on $X(\mathbf{C})$, hence constant; this constant drops out in the sum defining $\langle \mathfrak{x}, \mathfrak{y} \rangle_\infty$ because $\sum_i n_i = 0$, and similarly the choice of y_0 is irrelevant because $\sum_j m_j = 0$.)

Applying this general theory to Heegner divisors on $X_0(N)$, we find

$$\langle \mathfrak{x}_D, \mathfrak{x}_{D'} \rangle = \sum_{i=1}^h \sum_{i'=1}^{h'} G(\tau_{D,i}, \tau_{D',i'}) - \sum_p n(D, D', p) \log p \quad (7)$$

where G is an appropriate Green's function as above on $X_0(N)(\mathbf{C})$ (with $y_0 = \infty$) and the $n(D, D', p)$ are rational numbers which can be calculated explicitly using the theory of complex multiplication and our knowledge of a model of $X_0(N)$ over \mathbf{Z} . They turn out to be nonnegative and integral if $p \nmid N$ and to vanish

unless $p < DD'/4N$ (more precisely, unless $DD' = r^2 + 4Nmp$ for some integers r and $m > 0$).

What about G ? We need a function on $\mathfrak{H} \times \mathfrak{H}$ which is $\Gamma_0(N)$ -invariant and harmonic in each variable and has logarithmic singularities along the diagonal of $(\mathfrak{H}/\Gamma_0(N))^2$ and at infinity. A natural attempt is to set

$$G(z, z') = \sum_{\gamma \in \Gamma_0(N)/\pm 1} g(z, \gamma z') \quad (8)$$

where g is a function which is invariant under the diagonal action of $\Gamma_0(N)$, harmonic, and logarithmically singular on the diagonal of \mathfrak{H}^2 , and which drops off rapidly as the hyperbolic distance between z and z' tends to infinity. Such a function is

$$g(z, z') = \log \left| \frac{z - z'}{\bar{z} - \bar{z}'} \right|^2 \quad (z, z' \in \mathfrak{H}).$$

Unfortunately, it does not go to zero quite fast enough: the series (8) diverges like $\sum 1/n$. Instead, we replace the harmonicity condition $\Delta g = 0$, where

$$\Delta = \frac{1}{y^2} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$$

is the hyperbolic Laplacian, by $\Delta g = \varepsilon g$ with $\varepsilon > 0$, obtain a function g for which (8) does converge, and then let ε tend to zero. Specifically, set

$$g_s(z, z') = -2Q_{s-1} \left(1 + \frac{|z - z'|^2}{2yy'} \right) \quad (s \in \mathbf{C}, z, z' \in \mathfrak{H}),$$

where

$$Q_{s-1}(t) = \frac{\Gamma(s)^2}{2\Gamma(2s)} \left(\frac{1+t}{2} \right)^{-s} F \left(s, s; 2s; \frac{2}{1+t} \right) \quad (t > 1)$$

is the Legendre function of the second kind. It is invariant under the diagonal action of $\Gamma_0(N)$ (or even $\mathrm{SL}_2(\mathbf{R})$) on \mathfrak{H}^2 , has a logarithmic singularity on the diagonal, and (because of the second order ordinary differential equation satisfied by Q_{s-1}) satisfies $\Delta g_s = s(s-1)g_s$. The following series converges absolutely for $\mathrm{Re}(s) > 1$ and has the analogous properties on $(\mathfrak{H}/\Gamma_0(N))^2$:

$$G_s(z, z') = \sum_{\gamma \in \Gamma_0(N)/\pm 1} g_s(z, \gamma z'). \quad (9)$$

The desired Green's function G is obtained as

$$G(z, z') = \lim_{s \rightarrow 1} (G_s(z, z') - e_s(z, z'))$$

where $e_s(z, z')$ is a certain combination of Eisenstein series and elementary functions which we do not specify here.

Now that we have a formula for the Green's function we obtain its value at Heegner points as a certain explicit infinite sum of Legendre functions. Surprisingly, analytic techniques from the theory of modular forms (specifically: the "Rankin-Selberg method" when $D = D'$, and the theory of modular forms of

half-integral weight or the theory of Jacobi forms, combined with the theory of Hilbert modular forms on the real quadratic field $\mathbf{Q}(\sqrt{DD'})$, when $D \neq D'$ produce a different formula containing this same infinite sum. For $D \neq D'$ and N prime, for instance, one gets

$$\sum_{i=1}^h \sum_{i'=1}^{h'} G(\tau_{D,i}, \tau_{D',i'}) = \sum_p n(D, D', p) \log p + \frac{3}{4\pi} \sum_f \frac{c_f(D)c_f(D')}{\|g_f\|^2} L'(f, 1) \quad (10)$$

where, by a further miracle, the coefficients $n(D, D', p)$ are the same numbers as those occurring in (7) from the non-archimedean height computation. Together, of course, formulas (7) and (10) give a (highly nonconceptual) proof of Theorem 2, and similarly for Theorem 1.

It would, of course, be nice to have a more intrinsic proof of Theorems 1 and 2, directly involving the global heights. However, breaking down the identity into local pieces as in (10) does have the advantage of making more structure visible than in the smoother identities (3) and (5). We discuss two aspects of this.

First, consider the case $N = 1$. Here $X_0(N)$ has genus 0, so there are no cusp forms of weight 2 and the global height pairing is trivial (every degree zero divisor on \mathbf{P}^1 is principal). Hence both Theorem 1 and Theorem 2 are empty. But for this very reason (10) now becomes interesting. The Green's function for \mathbf{P}^1 is clearly $\log|x - y|^2$ (taking $y_0 = \infty$ in the definition (6)), so the Green's function for $X_0(1)$ is $\log|j(z) - j(z')|^2$, where

$$j(z) = e^{-2\pi iz} + 744 + 196884e^{2\pi iz} + \cdots \quad (z \in \mathfrak{H})$$

is the classical modular function giving the isomorphism from

$$X_0(1) = \mathfrak{H}/\mathrm{SL}_2(\mathbf{Z}) \cup \{\infty\}$$

to $\mathbf{P}^1(\mathbf{C})$. Hence (10) becomes

THEOREM 3 [8]. *Let D, D' be coprime discriminants of imaginary quadratic fields, h and h' their class numbers, and $\tau_{D,i}$ ($1 \leq i \leq h$), $\tau_{D',i'}$ ($1 \leq i' \leq h'$) the $\mathrm{SL}_2(\mathbf{Z})$ -inequivalent roots in H of quadratic equations of discriminants D, D' . Then*

$$\prod_{i=1}^h \prod_{i'=1}^{h'} |j(\tau_{D,i}) - j(\tau_{D',i'})|^2 = \prod_p p^{n(D, D', p)}$$

with explicitly given nonnegative integral exponents $n(D, D', p)$ which are nonzero only for primes p dividing one of the finitely many integers $(DD' - r^2)/4$, $|r| < \sqrt{DD'}$, $r \equiv DD' \pmod{2}$.

The point is that one knows by the classical theory of complex multiplication that the numbers $j(\tau_{D,i})$ and $j(\tau_{D',i'})$ are algebraic integers (in the Hilbert class fields of $\mathbf{Q}(\sqrt{D})$ and $\mathbf{Q}(\sqrt{D'})$, respectively), and Theorem 3 gives an explicit

formula for the absolute norm of their difference. As a numerical example, take $D = -163$ and $D' = -4$. Here $h = h' = 1$ and

$$\begin{aligned} j\left(\frac{1+i\sqrt{163}}{2}\right) - j(i) &= -262537412640768000 - 1728 \\ &= -2^6 3^6 7^2 11^2 19^2 127^2 163 \end{aligned}$$

in accordance with the theorem (the primes 11, 19, 127 and 163 divide $163 - n^2$ for $n = 3, 7, 6$, and 0, respectively).

Secondly, the fact that (10) is proved by purely analytic techniques from the theory of modular forms, rather than by height theory, means that one gets an analogous identity for forms of higher weight. It takes the form (for $N = 1$)

$$\begin{aligned} (DD')^{(k-1)/2} \sum_{i=1}^h \sum_{i'=1}^{h'} G_k(\tau_D, i, \tau_{D'}, i') \\ = \sum_p n_k(D, D', p) \log p + \frac{3\Gamma(k-1/2)}{2^{2k}\pi^{k+1/2}} \sum_f \frac{c_f(D)c_f(D')}{\|g_f\|^2} L'(f, k), \end{aligned} \quad (k = 3, 5, \dots) \quad (11)$$

where G_k is the resolvent kernel function defined by (9) (with $s = k$), the $n_k(D, D', p)$ are explicitly given integers which are zero unless p divides some positive integer of the form $(DD' - \tau^2)/4$, and the sum runs over all Hecke forms $f \in S_{2k}(\mathrm{SL}_2(\mathbf{Z}))$, the g_f being the forms of weight $k + \frac{1}{2}$ corresponding to the f and $c_f(D)$ the $|D|$ th Fourier coefficient of g_f . This identity suggests two problems:

(i) Find a "higher-weight height theory" which permits us to interpret (11) as a formula for $L'(f, k)$ in terms of heights. Motivated by the identity (11), Deligne and Brylinski have made some progress towards developing such a local height theory, the terms $G_k(\tau, \tau')$ and $n_k \log p$ in (11) appearing as the local contributions from the places ∞ and p to the height pairing of some higher weight Heegner cycles. However, there is as yet no global height theory, so one neither knows that the height pairing of a cycle with itself is nonnegative nor has a criterion for the vanishing of the global height.

(ii) In analogy with Theorem 3, prove that the individual numbers $G_k(\tau, \tau')$ in (11) are logarithms of algebraic numbers when $k = 3, 5$, or 7 (so that the sum over f is empty). Then (11) could be interpreted in these cases as giving the prime decomposition of the absolute norms of these algebraic numbers, just as for the j -values above. The conjectured algebraicity of $\exp(G_k(\tau, \tau'))$ for τ and τ' quadratic imaginary numbers would seem to be an interesting property of the resolvent kernel function, since it shows that G_k is a new transcendental function whose special values can be used to give algebraic extensions, in the spirit of complex multiplication theory and Kronecker's Jugendtraum. It is supported by the analogy with the case $k = 1$, by the fact that (11) provides a proof whenever $h = h' = 1$, and by numerical evidence. (A numerical example for $k = 2$, $D = D' = -23$ is given at the end of [9].) The restriction to the handful of cases

with $S_{2k}(N) = \{0\}$ can be circumvented by replacing the G_k by appropriate linear combinations of their images under Hecke operators.

BIBLIOGRAPHY

1. B. J. Birch, *Heegner points of elliptic curves*, Symposia Mathematica, Vol. 15, Academic Press, London, 1975, pp. 441–445.
2. M. Eichler and D. Zagier, *The theory of Jacobi forms*, Progr. Math., vol. 55, Birkhäuser, Boston, Mass., 1985.
3. D. Goldfeld, *The class numbers of quadratic fields and the conjectures of Birch and Swinnerton-Dyer*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. **3** (1976), 623–663.
4. B. Gross, *Heegner points on $X_0(N)$* , Modular Forms (R. A. Rankin, ed.), Ellis Horwood, Chichester, 1984, pp. 87–106.
5. —, *Local heights on curves*, Arithmetic Geometry (G. Cornell, J. H. Silverman, eds.), Springer-Verlag, Berlin-Heidelberg-New York, 1986, pp. 327–339.
6. B. Gross, W. Kohnen, and D. Zagier, *Heegner points and derivatives of L -series*. II, Math. Ann. (to appear).
7. B. Gross and D. Zagier, *Points de Heegner et dérivées de fonctions L* , C. R. Acad. Sci. Paris Ser. I Math. **297** (1983), 85–87.
8. —, *On singular moduli*, J. Reine Angew. Math. **355** (1985), 191–220.
9. —, *Heegner points and derivatives of L -series*, Invent. Math. **84** (1986), 225–320.
10. W. Kohnen, *Fourier coefficients of modular forms of half-integral weight*, Math. Ann. **271** (1985), 237–268.
11. W. Kohnen and D. Zagier, *Values of L -series of modular forms at the center of the critical strip*, Invent. Math. **64** (1981), 175–198.
12. J. Oesterlé, *Nombres de classes des corps quadratiques imaginaires*, Séminaire Bourbaki, Vol. 1983/84, Astérisque No. 121–122 (1985), 309–323.
13. G. Shimura, *On modular forms of half-integral weight*, Ann. Math. **97** (1973), 440–481.
14. M.-F. Vignéras, *Valeur au centre de symétrie des fonctions L associées aux formes modulaires*, Séminaire de Théorie des Nombres, Paris 1979–1980, Progr. Math., vol. 12, Birkhäuser, Boston, Mass., 1981, pp. 331–356.
15. J.-L. Waldspurger, *Sur les coefficients de Fourier des formes modulaires de poids demi-entier*, J. Math. Pures Appl. **60** (1981), 375–484.
16. —, *Correspondances de Shimura et quaternions*, Preprint.

UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND 20742, USA

MAX-PLANCK-INSTITUT FÜR MATHEMATIK, D-5300 BONN 3, FEDERAL REPUBLIC OF GERMANY

Inner Functions: Results, Methods, Problems

A. B. ALEKSANDROV

1. Introduction. A function f , holomorphic in the open unit ball $B = B^n$ of \mathbb{C}^n , is said to be inner if $|f| \leq 1$ everywhere on B and $|\lim_{r \rightarrow 1-} f(r\zeta)| = 1$ almost everywhere on the unit sphere $S = S^n = \partial B^n$. The existence of nonconstant inner functions in the ball B^n ($n \geq 2$) was proved independently by the author [2] and E. Løw [20] about five years ago. In the one-dimensional case (see [19]) the inner functions play an important role in the function theory on the unit disc $\mathbf{D} \stackrel{\text{def}}{=} B^1$. See [25, 26, 27, 29] for the existence problem for inner functions in the ball and related questions. The author [2] used an L^p -technique for $p < 1$ whereas E. Løw [20] modified a construction of M. Hakim and N. Sibony [15]. Afterwards it was observed (see [5]) that both methods of construction of inner functions are in a sense equivalent. Later, a simpler construction of nonconstant inner functions appeared in [5] and [31]. These constructions are based on the very important and beautiful result of J. Ryll and P. Wojtaszczyk [28].

We state three theorems that refine and generalize the theorem on the existence of nonconstant inner functions in the ball B .

Let φ be a positive continuous function on the closed ball \overline{B} . Denote by σ the probability rotation-invariant measure on S .

THEOREM A. *There exists a function f holomorphic on B such that*

- (1) $|f| < \varphi$ everywhere on B ,
- (2) $|f^*| = \varphi$ almost everywhere on S .

Here f^* denotes the boundary values of f .

THEOREM B. *Let $\varepsilon > 0$. Then there exists a closed set $K \subset S$ and a function f continuous on $\overline{B} \setminus K$ such that*

- (1) $f|_B$ is holomorphic,
- (2) $\sigma(K) = 0$,
- (3) $|f| < \varphi$ everywhere on B ,
- (4) $|f| > \varphi - \varepsilon$ everywhere on $S \setminus K$.

THEOREM C. *Let $\varepsilon > 0$. Then there exists a function f continuous on \overline{B} such that*

- (1) $f|_B$ is holomorphic,

- (2) $|f| < \varphi$ everywhere on B ,
 (3) $\sigma\{\zeta \in S : |f(\zeta)| \neq \varphi(\zeta)\} < \varepsilon$.

Theorem A was proved by the author [2] and E. Løw [20] (see also [5, 21]), Theorem B by M. Hakim and N. Sibony [15], and Theorem C by the author [4]. We have not stated the above theorems in full generality. For example, in Theorems A and C it is sufficient to assume that φ is lower semicontinuous rather than continuous. In Theorem B one can in addition require f to be holomorphic on $\overline{B} \setminus K$ (see [14]). E. Løw [21] extended Theorems A, B, C to the strictly pseudoconvex domains with C^2 boundary and to the weakly pseudoconvex domains with C^∞ -boundary. The author [5] extended Theorems A, B, C to the Siegel domains of I and II type and, in particular, to the classical domains.

Theorems A, B, C hold for some nonsmooth pseudoconvex domains with a small set of singular points on the boundary. For example, they hold for

$$B = \left\{ z \in \mathbb{C}^n : \sum_{j=1}^n |z_j|^{p_j} < 1 \right\},$$

where p_j are positive numbers.

The methods of the papers [15, 20, and 2] allowed one to obtain results concerning interpolation, approximation, zero-varieties of holomorphic functions, and also some other results (see [2-6, 8, 14, 16, 17, 21-23, 27, 30, 32]).

B. Tomaszewski [32] proved that in Theorem C one can require in addition that $f \in \text{Lip } \alpha$ for some $\alpha > 0$ if $\varphi \in \text{Lip } 1$. Thus there exist an $\alpha > 0$ and a function f , holomorphic in B and continuous up to the boundary such that $f \in \text{Lip } \alpha$, $|f| < 1$ everywhere in B and $m\{\zeta \in S : |f(\zeta)| = 1\} > 0$. It is known that for $\alpha > \frac{1}{2}$ such a function f does not exist, see [29, 26, 18].

For more recent results one should note the proof of the existence of proper holomorphic maps from the ball into a polydisc of sufficiently large dimension (E. Løw [22] and the author [8]) and from every strictly pseudoconvex domain with C^2 boundary into a ball of sufficiently large dimension (E. Løw [23]). Moreover the following two theorems are valid.

THEOREM A' [22, 8]. *If $N \geq \kappa = \kappa(n)$, then there exists a holomorphic map $f = (f_1, f_2, \dots, f_N) : B^n \rightarrow \mathbb{C}^N$ such that*

- (1) $|f_j| < \varphi$ everywhere on B^n ($1 \leq j \leq n$), and
 (2) $\lim_{z \rightarrow \zeta} \max_{1 \leq j \leq N} |f_j(z)| = \varphi(\zeta)$ for all $\zeta \in S$.

For $\varphi \equiv 1$, Theorem A' gives a proper holomorphic map from the ball B^n into the polydisc \mathbb{D}^N . Theorem A' remains valid if we replace B^n by any strictly pseudoconvex domain with C^2 boundary (see [23, 8]). This follows also from Theorem A'' stated below.

THEOREM A'' [23]. *Let $\Omega \subset \mathbb{C}^n$ be a strictly pseudoconvex domain with C^2 boundary. Then for $N \geq \kappa = \kappa(n)$ there exists a continuous map $f = (f_1, f_2, \dots, f_N) : \overline{\Omega} \rightarrow \mathbb{C}^N$ such that*

- (1) $f|_\Omega$ is holomorphic,

- (2) $\sum_{j=1}^N |f_j|^2 < \varphi^2$ everywhere on Ω , and
 (3) $\sum_{j=1}^N |f_j|^2 = \varphi^2$ everywhere on $\partial\Omega$.

For $\varphi \equiv 1$, Theorem A'' gives a proper holomorphic map from Ω into B^N continuous up to the boundary $\partial\Omega$.

2. Problems concerning inner functions in the ball. Every bounded holomorphic function f on B has naturally defined boundary values not only σ -almost everywhere on S but also μ -almost everywhere on S for arbitrary Henkin measure μ on S (see [26]).

Problem 1. Does there exist a nonconstant inner function f on B^n ($n \geq 2$) such that $|f^*| = 1$ μ -almost everywhere for all Henkin measures μ ?

The author does not know whether there is a nonconstant inner function f on B^n ($n \geq 2$) such that for all $\zeta \in S$ the slice function f_ζ defined by $f_\zeta(z) = f(z\zeta)$ is an inner function on \mathbf{D} ?¹

A positive solution of Problem 1 would imply a positive answer to this question.

We say that a regular Borel measure μ on S is a P -measure if its Poisson integral is pluriharmonic (the real part of a holomorphic function) in B . For any nonconstant inner function f on B there exists a positive singular P -measure μ on S such that $\operatorname{Re}((1+f)/(1-f))$ is the Poisson integral of μ .

It is known (see [11, 12, 7]) that a set supporting a P -measure cannot be small. It would be interesting to know what Hausdorff dimension such a set can have.

In many questions of the function theory in the ball it is useful to consider the quasimetric d on S ($d(z, \zeta) \stackrel{\text{def}}{=} |1 - \langle z, \zeta \rangle|$, where $\langle \cdot, \cdot \rangle$ denotes the scalar product) instead of the usual Euclid metric. If $n \geq 2$, the triangle inequality is not true for d . However \sqrt{d} is a metric on S (see [26]); therefore

$$d(\zeta, \xi) \leq 2(d(\zeta, z) + d(z, \xi))$$

for all $\zeta, \xi, z \in S$. To this quasimetric d we associate (in a usual way) the α -dimensional Hausdorff measure h_α . We call h_α a complex Hausdorff measure on S of dimension α . The family of measures $\{h_\alpha\}$ allows us to define the Hausdorff dimension, which we denote by CH-dim and call the complex Hausdorff dimension. It is not difficult to verify that CH-dim (S^n) = n .

Set

$$Q(\zeta, r) \stackrel{\text{def}}{=} \{z \in S : d(z, \zeta) < r\}.$$

It is easy to prove (see [7]) the inequality

$$\mu(Q(\zeta, r)) \leq C(n)r^{n-1}||\mu||$$

for all positive P -measures μ on S , all $\zeta \in S$, and all $r > 0$. This inequality shows that if $h_{n-1}(E) = 0$, then $\mu(E) = 0$ for every positive P -measure μ . In

¹Professor W. Rudin has communicated to me that Y. Dupain obtained a positive answer to this question.

particular, if the Borel set $E \subset S$ supports a positive P -measure μ ($\mu \neq 0$), then $h_{n-1}(E) > 0$. Hence $\text{CH-dim } E \geq n - 1$. The author does not know whether the equality $\text{CH-dim } E = n - 1$ may occur.

Problem 2. For which $\alpha \in [n - 1, n)$ does there exist a positive singular P -measure μ ($\mu \neq 0$) on S supported by a set of finite (or σ -finite) h_α -measure?

Problem 3. Are there a Borel set $E \subset S$ and $\alpha \in [n - 1, n)$ such that $0 < h_\alpha(E) < +\infty$ and $1_E \cdot h_\alpha$ is a P -measure? (1_E denotes the characteristic function of the set E .) For $\alpha = n$ the answer to the last question is positive. Namely, for every $t \in [0, h_n(S)]$ there exists a Borel set $E \subset S$ such that $h_n(E) = t$ and $1_E \cdot \sigma$ is a P -measure (see [25]).

To every positive P -measure μ on S we associate its Poisson integral, a pluriharmonic function u on B . The slice function u_ζ (defined by $u_\zeta(z) = u(\zeta z)$) is a positive harmonic function on D . Hence u_ζ is the Poisson integral of a positive measure μ_ζ on $\mathbf{T} \stackrel{\text{def}}{=} S^1 = \partial D$.

Problem 4. Can the measure μ_ζ be atomic for almost all $\zeta \in S$?

This problem is connected with Problem 2 and with the following question.

Does there exist a nonconstant inner function on B^n ($n \geq 2$) such that for almost all $\zeta \in S$ the slice function f_ζ has a finite angular derivative (see [10]) almost everywhere on \mathbf{T} ?

3. Inner function on uniform algebras. First we briefly describe the abstract approach to Theorems A, B, C (see [5]). Let K be a compact Hausdorff space. We denote by $C(K)$ the space of all (complex) continuous functions on K . Let $M(K)$ denote the space of all (complex) regular Borel measures on K . Let A be a closed subspace of $C(K)$ and μ a positive measure in $M(K)$. In [5], the concept of a regular triple (A, K, μ) is introduced. From several equivalent definitions of the regular triple (see [5]) we choose the following one. A triple (A, K, μ) is said to be regular if there exists a number $\tau > 0$ such that the inequality

$$\sup \left\{ \int_K |f|^2 d\mu : f \in A, |f| < \varphi \right\} \geq \tau \int_K \varphi^2 d\mu \quad (1)$$

holds for all positive functions $\varphi \in C(K)$. In this case we write $(A, K, \mu) \in (*)$. If $(A, K, \mu) \in (*)$, then the equality

$$\sup \left\{ \int_K |f|^2 d\mu : f \in A, |f| < \varphi \right\} = \int_K \varphi^2 d\mu$$

holds, in fact, for all positive functions $\varphi \in C(K)$ (see [5]). The results of the paper [5] allow one to reduce Theorems A, B, C to the regularity of the triple (A, K, μ) , where $K = \overline{B}$, $\mu = \sigma$, and A is the ball algebra $A(B)$, i.e., the space of all holomorphic functions in B continuous up to S .

Suppose that the space A contains the constants and separates the points of K . Then we may consider the Choquet boundary P_A (see [24] for the definition). It is known that $\partial_A \stackrel{\text{def}}{=} \overline{P_A}$ is the Shilov boundary of the space A (see [24]). One can prove that if $(A, K, \mu) \in (*)$, then $\text{supp } \mu \subset \partial_A$. Moreover the following theorem holds.

THEOREM 1. *If $(A, K, \mu) \in (*)$, then $\mu(\Delta) = 0$ for every Baire set $\Delta \subset K \setminus P_A$.*

We shall show that the converse is false in general, although it holds for atomic measures μ . If K is a metrizable compact, then the Choquet boundary is a Baire set. Hence in this case $(A, K, \mu) \in (*)$ implies $\mu(K \setminus P_A) = 0$.

Now let us consider more closely the case when A is a uniform algebra. The following theorem shows that usually to prove the regularity of (A, K, μ) it is sufficient to verify the regularity of $(A|_{\partial_A}, \partial_A, \mu)$. Here we suppose that $\text{supp } \mu \subset \partial_A$ since otherwise we would have $(A, K, \mu) \notin (*)$ by Theorem 1.

THEOREM 2. *Let A be a uniform algebra on K , and let μ be a positive measure in $M(\partial_A)$. Suppose that for every point $x \in K \setminus \partial_A$ there exists a neighborhood of x intersecting only a finite number of Gleason parts of A . Then $(A|_{\partial_A}, \partial_A, \mu) \in (*)$ implies $(A, K, \mu) \in (*)$.*

The regularity of the triple $(A(B), \overline{B}, \mu)$ is proved in [5] for every positive measure $\mu \in M(S)$. This implies the following theorem.

THEOREM 3. *Let A be a uniform algebra on K . Suppose that there exists a finite sequence $\{f_j\}_{j=1}^N$ of functions in A separating the points of K and such that $\sum_{j=1}^N |f_j|^2 = 1$ everywhere on ∂_A . Then $(A, K, \mu) \in (*)$.*

Theorem 3 and Theorem A'' imply the regularity of the triple $(A(\Omega), \overline{\Omega}, \mu)$ for every positive measure $\mu \in M(\partial\Omega)$, where Ω is a strictly pseudoconvex domain of \mathbb{C}^n with C^2 boundary and $A(\Omega)$ is the space of all $f \in C(\overline{\Omega})$ holomorphic on Ω .

The author does not know whether one can consider an infinite sequence $\{f_j\}_{j=1}^\infty$ in Theorem 3. This question is connected with Problem 5 stated below.

Now we pass to particular uniform algebras. For each compact K of \mathbb{C} we consider the following three algebras:

- (1) $P(K)$, the closed subalgebra of $C(K)$ generated by 1 and z ;
- (2) $R(K)$, the closed subalgebra of $C(K)$ generated by the family of functions $\{(z - a)^{-1}\}$, $a \in \mathbb{C} \setminus K$;
- (3) $A(K)$, the algebra of all functions $f \in C(K)$ holomorphic on the interior of K .

Denote by \hat{K} the polynomial convex hull of K . It is well known that $\partial\hat{K}$ is the Shilov boundary of $P(K)$ and that $P(K)$ is the Dirichlet algebra. This assertion and Theorem 1 imply the following theorem.

THEOREM 4. *The triple $(P(K), K, \mu)$ is regular if and only if $\text{supp } \mu \subset \partial\hat{K}$.*

For the algebras $R(K)$ and $A(K)$ the situation is more difficult. Let $\{\Omega_j\}_{j \in J}$ be the family of all connected components of $\hat{\mathbb{C}} \setminus K$. Set $\partial'K \stackrel{\text{def}}{=} \bigcup_{j \in J} \partial\Omega_j$. It is clear that $\partial'K$ is contained in the Choquet boundary of $R(K)$ and $A(K)$.

Theorem 4 implies the following:

THEOREM 5. *Let μ be a positive measure in $M(K)$. Then if $\mu(K \setminus \partial' K) = 0$, then $(R(K), K, \mu) \in (*)$ and $(A(K), K, \mu) \in (*)$.*

The converse is false in general. Theorem 4 says that the converse of Theorem 1 with $A = P(K)$ holds. We show that the converse of Theorem 1 with $A = R(K)$ or $A = A(K)$ is not true in general.

We give an example of a compact subset K of \mathbf{C} and a positive measure μ on K such that $R(K) = A(K)$, $\text{supp } \mu \subset \partial K$, ∂K is the Choquet boundary of $R(K) = A(K)$ but $(R(K), K, \mu) \notin (*)$.

For K we take the "string of beads" (see [13]). Then $R(K)$ is a standard example of an algebra with a nonconnected Gleason part (see [13]). The compact set K is constructed by deleting a sequence of pairwise disjoint discs with centers on the real axis from the closed unit disc so that

- (1) $K \cap \mathbf{R}$ has a positive line measure.
- (2) $K \cap \mathbf{R}$ has no interior points with respect to \mathbf{R} , i.e., $K \cap \mathbf{R} = (\partial K) \cap \mathbf{R}$.
- (3) Each point in $K \cap \mathbf{R}$ is a peak point for the algebra $R(K)$, i.e., ∂K is the Choquet boundary of $R(K)$.

It is known that $R(K) = A(K)$ for this compact set K (see [13]).

Let μ be the one-dimensional Hausdorff measure on ∂K . To prove $(R(K), K, \mu) \notin (*)$ we use the following lemma.

LEMMA. *Let f be a bounded holomorphic function on $\text{int}(K)$. Suppose that*

$$\lim_{y \rightarrow 0+} f(x + iy) = \lim_{y \rightarrow 0-} f(x + iy)$$

for μ -almost all $x \in K \cap \mathbf{R}$. Set $f^(x) = \lim_{y \rightarrow 0} f(x + iy)$. Then $\mu\{x \in \mathbf{R} \cap K : |f^*(x)| = 1\} > 0$ implies $f(z)f(\bar{z}) = 1$ for all $z \in \text{int } K$.*

To prove this lemma it is enough to apply the boundary uniqueness theorem (see [19]).

THEOREM 6. *The triple $(R(K), K, \mu)$ is not regular.*

PROOF. Suppose that $(R(K), K, \mu) \in (*)$. Then the results of [5] allow us to construct a function f holomorphic on $\text{int}(K)$ such that $|f(z)| < 1$ for all $z \in \text{int}(K)$ and $\lim_{y \rightarrow 0+} f(x + iy) = \lim_{y \rightarrow 0-} f(x + iy) \in \mathbf{T}$ for μ -almost all $x \in \mathbf{R} \cap K$. Hence, by the Lemma, $1 = f(z)f(\bar{z}) < 1$ and we get a contradiction. Q.E.D.

Now let K be a convex compact subset of a locally convex Hausdorff space X . By $P(K)$ we denote the uniform algebra on K generated by the family of functions $\{f|_K\}_{f \in X^*}$, where X^* is the dual space. By $\text{b}K$ we denote the Choquet boundary of the algebra $P(K)$. In [9] a simple geometric description of $\text{b}K$ is presented.

Problem 5. Let μ be a positive measure in $M(K)$ such that $\mu(\Delta) = 0$ for every Baire set $\Delta \subset K \setminus \text{b}K$. Is the triple $(P(K), K, \mu)$ regular?

If K is metrizable, then this question can be restated as follows.

Problem 5'. Let μ be a positive measure in $M(K)$ supported by bK . Is the triple $(P(K), K, \mu)$ regular?

I do not know an answer to this question even if $K \subset \mathbf{C}^n$ ($n \geq 2$).

It would be interesting to consider the case when K is the closed unit ball of the Hilbert space. Set

$$l^2 \stackrel{\text{def}}{=} \left\{ x = \{x_n\}_{n=1}^\infty : \|x\|_2 \stackrel{\text{def}}{=} \left(\sum_{n=1}^\infty |x_n|^2 \right)^{1/2} < +\infty \right\},$$

$$B^\infty \stackrel{\text{def}}{=} \{x \in l^2 : \|x\|_2 < 1\}, \quad S^\infty \stackrel{\text{def}}{=} \{x \in l^2 : \|x\|_2 = 1\}, \quad K = B^\infty \cup S^\infty.$$

The set K is compact in the weak topology of the Hilbert space l^2 . In this particular case Problem 5 is restated as follows. Let μ be a positive measure in $M(\overline{B}^\infty)$ supported by S^∞ . Is the triple $(P(\overline{B}^\infty), \overline{B}^\infty, \mu)$ regular?

It seems possible that this question is connected with the following problem.

Set $A(r, n) \stackrel{\text{def}}{=} \sup \int_{S^n} |f(r\zeta)|^2 d\sigma(\zeta)$, where the supremum is taken over all functions f holomorphic in B^n such that $f(0) = 0$ and $|f| < 1$ everywhere on B^n .

Problem 6. Compute or estimate $A(r, n)$ ($0 < r < 1$, $n = 2, 3, \dots$).

4. Harmonic vector fields. In [3] it is shown that methods of M. Hakim, N. Sibony, and E. Løw [15, 20] enable one to obtain an analog of Theorems A and B for the gradient of a harmonic function. Here we state only an analog of Theorem A. In this section $B = B^n$ denotes the open unit ball of \mathbf{R}^n , $S = \partial B$. Let φ be a positive continuous function on \overline{B} ; $0 < p \leq +\infty$; $n \geq 2$.

THEOREM A_h. *There exists a real harmonic function u on B^n such that $|\nabla u|_p < \varphi$ everywhere on B^n and $|(\nabla u)^*|_p = \varphi$ almost everywhere on S^n . Here*

$$|\nabla u|_p \stackrel{\text{def}}{=} \left(\sum_{j=1}^n \left| \frac{\partial u}{\partial x_j} \right|^p \right)^{1/p}, \quad |(\nabla u)^*(\zeta)|_p \stackrel{\text{def}}{=} \lim_{r \rightarrow 1-} |(\nabla u)(r\zeta)|_p.$$

Theorem A_h for $n = 2$ has the following corollary.

COROLLARY. *There exists a bounded holomorphic function on \mathbf{D} such that $|f^*|_p = \varphi$ almost everywhere on \mathbf{T} , where*

$$|f^*(\zeta)|_p \stackrel{\text{def}}{=} \lim_{r \rightarrow 1-} (|\operatorname{Re} f(r\zeta)|^p + |\operatorname{Im} f(r\zeta)|^p)^{1/p}.$$

For $p = 2$ this corollary is well known. In this case we may consider an arbitrary nonnegative function $\varphi \in L^\infty(\mathbf{T})$. Then the existence of a bounded holomorphic function $f \not\equiv 0$ on \mathbf{D} such that $|f^*|_2 = \varphi$ is equivalent to the following condition: $\log \varphi \in L^1(\mathbf{T})$. Moreover, in this case, there exists an outer function f (see [19]). I do not know whether f can be taken outer in the Corollary when $p \neq 2$.

Problem 7. Let φ be a nonnegative function in $L^\infty(\mathbf{T})$ and let $p \in (0, 2) \cup (2, +\infty)$. Suppose that $\log \varphi \in L^1(\mathbf{T})$. Does there exist a bounded holomorphic function f on \mathbf{D} such that $|f^*|_p = \varphi$ almost everywhere on \mathbf{T} ?

The answer is "yes" if φ is positive and lower semicontinuous. For such φ , Theorem A_B remains valid.

ACKNOWLEDGMENT. I wish to thank N. G. Makarov for assistance in the preparation of the English version of this paper.

REFERENCES

1. A. B. Aleksandrov, *Essays on nonlocally convex Hardy classes*, Lecture Notes in Math., vol. 864, Springer-Verlag, 1981, pp. 1–89.
2. —, *The existence of inner functions in the ball*, Mat. Sb. **118** (160) (1982), 147–163; English transl. in Math. USSR-Sb. **46** (1983).
3. —, *Inner functions on spaces of homogeneous type*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. LOMI **126** (1983), 7–14. (Russian)
4. —, *On boundary values of functions holomorphic in a ball*, Dokl. Akad. Nauk SSSR **271** (1983), 777–779; English transl. in Soviet Math. Dokl. **28** (1983).
5. —, *Inner functions on compact spaces*, Funktsional. Anal. i Prilozhen. **18** (1984), 1–13. (Russian)
6. —, *The Blaschke condition and roots of bounded holomorphic functions*, Multivariate Complex Analysis, Krasnoyarsk, 1985, pp. 23–26. (Russian)
7. —, *Function theory in the ball*, deposited at VINITI, 1985, 115–190. (Russian)
8. —, *Proper holomorphic maps from the ball in a polydisc*, Dokl. Akad. Nauk SSSR **286** (1986), 11–15. (Russian)
9. E. L. Arenson, *The Gleason parts and the Choquet boundary of a function algebra on a convex compactum*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **113** (1981), 202–207. (Russian)
10. C. Carathéodory, *Funktionentheorie*, Vols. 1–2, Birkhäuser, Basel, 1950.
11. F. Forelli, *Measures whose Poisson integrals are pluriharmonic*, Illinois J. Math. **18** (1974), 373–388.
12. —, *Measures whose Poisson integrals are pluriharmonic. II*, Illinois J. Math. **19** (1975), 584–592.
13. T. W. Gamelin, *Uniform algebras*, Prentice-Hall, Englewood Cliffs, N.J., 1969.
14. M. Hakim, *Valeurs au bord de fonctions holomorphes bornées en plusieurs variables complexes*, Séminaire Bourbaki, 1982/83, Astérisque no. 105–106, pp. 293–305.
15. M. Hakim and N. Sibony, *Fonctions holomorphes bornées sur la boule unité de \mathbf{C}^n* , Invent. Math. **67** (1982), 213–222.
16. —, *Ensemble des zéros d'une fonction holomorphe bornée dans la boule unité*, Math. Ann. **280** (1982), 469–474.
17. —, *Valeurs au bord des modules de fonctions holomorphes*, Math. Ann. **264** (1983), 197–210.
18. G. M. Henkin and J. Leiterer, *Theory of functions on complex manifolds*, Akademie-Verlag, Berlin, 1984.
19. P. Koosis, *Lectures on H_p spaces*, London Math. Soc. Lecture Note Ser., vol. 40, Cambridge Univ. Press, 1980.
20. E. Løw, *A construction of inner functions on the unit ball in \mathbf{C}^p* , Invent. Math. **67** (1982), 223–229.
21. —, *Inner functions and boundary values in $H^\infty(\Omega)$ and $A(\Omega)$ in smoothly bounded pseudoconvex domains*, Math. Z. **185** (1984), 191–210.
22. —, *The ball in \mathbf{C}^n is a closed complex submanifold of a polydisc*, Preprint.
23. —, *Embeddings and proper holomorphic maps of strictly pseudoconvex domains into polydiscs and balls*, Preprint.
24. R. R. Phelps, *Lectures on Choquet's theorem*, Van Nostrand Math. Studies No. 7, Van Nostrand, Princeton, N.J., 1966.

25. W. Rudin, *The inner function problem in balls*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **81** (1978), 278–280.
26. —, *Function theory in the unit ball of \mathbb{C}^n* , Grundlehren. Math. Wiss., vol. 241, Springer-Verlag, 1980.
27. —, *Inner functions in the unit ball of \mathbb{C}^n* , J. Funct. Anal. **50** (1983), 100–126.
28. J. Ryll and P. Wojtaszczyk, *On homogeneous polynomials on a complex ball*, Trans. Amer. Math. Soc. **276** (1983), 107–116.
29. N. Sibony, *Valeurs au bord de fonctions holomorphes et ensembles polynomialement convexes*, Lecture Notes in Math., vol. 578, Springer-Verlag, 1977, pp. 300–313.
30. B. Tomaszewski, *Interpolation and inner maps preserve measure*, J. Funct. Anal. **55** (1984), 63–67.
31. —, *A simple proof of the existence of inner functions on the unit ball of \mathbb{C}^d* , Preprint.
32. —, *Interpolation by Lipschitz holomorphic functions*, Preprint.

LENINGRAD STATE UNIVERSITY, LENINGRAD, USSR

Regularity of Solutions of the $\bar{\partial}$ -Neumann Problem

DAVID CATLIN

Let f be a $\bar{\partial}$ -closed $(0,1)$ -form on a smoothly bounded domain Ω in \mathbb{C}^n . The $\bar{\partial}$ -Neumann problem is a nonelliptic boundary value problem that arises when one tries to find a particular solution u of the equation $\bar{\partial}u = f$, namely, the solution u that is orthogonal to the subspace of holomorphic functions in $L^2(\Omega)$. To state this problem precisely, note that if we interpret $\bar{\partial}$ in the sense of distribution theory, then we obtain

$$L^2(\Omega) \xrightarrow{\bar{\partial}} L^2_{(0,1)}(\Omega) \xrightarrow{\bar{\partial}} L^2_{(0,2)}(\Omega).$$

For each $\bar{\partial}$ -operator we may form its L^2 -adjoint:

$$L^2(\Omega) \xleftarrow{\bar{\partial}^*} L^2_{(0,1)}(\Omega) \xleftarrow{\bar{\partial}^*} L^2_{(0,2)}(\Omega).$$

In analogy with Hodge theory, we define an operator on $(0,1)$ -forms U by

$$\square U = (\bar{\partial}\bar{\partial}^* + \bar{\partial}^*\bar{\partial})U.$$

The importance of this operator lies in the fact that if f is a $\bar{\partial}$ -closed $(0,1)$ -form and $\square U = f$, then $\bar{\partial}U = 0$. Thus $u = \bar{\partial}^*U$ satisfies $\bar{\partial}u = f$. Moreover, u is orthogonal to the holomorphic functions, since it is in the range of the adjoint. We will call u the Kohn solution.

The operator \square is actually a differential operator, for if $v \in \text{Dom}(\bar{\partial}^*)$, then $\bar{\partial}^*v = \vartheta v$, where ϑ is the formal adjoint of $\bar{\partial}$. For $\square U$ to be well defined, we need that

$$U \in \text{Dom}(\bar{\partial}^*), \tag{1}$$

$$\bar{\partial}U \in \text{Dom}(\bar{\partial}^*), \tag{2}$$

even when $U \in C^\infty_{(0,1)}(\bar{\Omega})$. The first condition amounts to a Dirichlet condition on one of the components of U , whereas the second one is equivalent to a Neumann condition in an antiholomorphic direction that is transverse to $b\Omega$. The $\bar{\partial}$ -Neumann problem then is the problem of finding a solution U of $\square U = f$ satisfying the boundary conditions given by (1) and (2). It was proposed by Spencer and formulated precisely as above by Kohn [10]. It has been applied to

a number of analytic questions concerning the boundary behavior of holomorphic functions. Among the most striking is the following theorem of Bell and Ligocka [1] which generalizes a theorem of Fefferman [8]:

THEOREM [1]. *Let Ω_i , $i = 1, 2$, be smoothly bounded domains in \mathbb{C}^n and suppose that f is a biholomorphism from Ω_1 to Ω_2 . Suppose further that each domain satisfies the following condition: if α is a $\bar{\partial}$ -closed form in $C_{(0,1)}^\infty(\bar{\Omega}_i)$, then the Kohn solution u of $\bar{\partial}u = \alpha$ satisfies $u \in C^\infty(\bar{\Omega}_i)$, $i = 1, 2$. Then the map f extends to a smooth diffeomorphism of $\bar{\Omega}_1$ onto $\bar{\Omega}_2$.*

In order to prove that the Kohn solution is in $C^\infty(\bar{\Omega})$, it suffices to prove that $U \in C_{(0,1)}^\infty(\bar{\Omega})$, for $u = \bar{\partial}^*U = \partial U$. Since \square is elliptic in the interior of Ω , this presents a problem only near the boundary. One of the principal methods of proving boundary regularity is the method of a priori subelliptic estimates. In a small neighborhood V about a boundary point z_0 , we may choose real coordinates $(x_1, \dots, x_{2n-1}, r)$, where $\Omega = \{z; r(z) < 0\}$. We define a tangential Fourier transform by

$$\tilde{\varphi}(\xi, r) = \int \varphi(x', r) e^{-i\langle x', \xi \rangle} dx', \quad \xi \in \mathbb{R}^{2n-1}, \varphi \in C_0^\infty(V),$$

and a tangential Sobolev norm by

$$|||\varphi|||_\varepsilon^2 = \int_{-\infty}^0 \int_{\mathbb{R}^{2n-1}} |\tilde{\varphi}(\xi, r)|^2 (1 + |\xi|^2)^\varepsilon d\xi dr.$$

Let $Q(U, U) = \|\bar{\partial}U\|^2 + \|\bar{\partial}^*U\|^2$. A subelliptic estimate of order $\varepsilon > 0$ is said to hold in V if

$$|||U|||_\varepsilon^2 \leq C(Q(U, U) + \|U\|^2) \quad (3)$$

holds for all $(0, 1)$ -forms U in $\text{Dom}(\bar{\partial}) \cap \text{Dom}(\bar{\partial}^*)$ that are supported in V . Kohn and Nirenberg [13] have shown that such an estimate in V implies local regularity of U :

THEOREM [13]. *Suppose that $\square U = f$ and that $f|_V \in C_{(0,1)}^\infty(V)$. If (3) holds in V , then $U|_V \in C_{(0,1)}^\infty(V)$.*

When the boundary of Ω is strongly pseudoconvex at z_0 , then Kohn showed that (3) holds with $\varepsilon = 1/2$. In order to describe further results in the weakly pseudoconvex case, we must introduce the notion of type. Following D'Angelo [6], suppose that a one-dimensional complex-analytic variety X (possibly singular at z_0) contains z_0 and is parametrized by

$$X = \{\gamma(s) = (\gamma_1(s), \dots, \gamma_n(s)), s \in \mathbb{C}, |s| < \delta\}$$

with $\gamma(0) = z_0$. If g is a smooth function near the origin, let $\nu(g)$ denote the order of vanishing of g at the origin. If g is vector-valued, then set $\nu(g) = \min_{i=1, \dots, N} \nu(g_i)$. Now define

$$\tau(x; z_0) = \nu(r \circ \gamma) / \nu(\gamma).$$

$\tau(x; z_0)$ is clearly independent of the parametrization γ , and it represents the order to which the variety X passes out of $b\Omega$ at z_0 . Now set

$$T(z_0) = \sup_X \{\tau(X; z_0)\}.$$

Surprisingly, it turns out that $T(z)$ is not an upper semicontinuous function of z in $b\Omega$ [6]. However, by a theorem of D'Angelo [6], it is locally finite:

THEOREM [6]. *Suppose that $T(z_0) < \infty$. Then for all z sufficiently close to z_0 ,*

$$T(z) \leq (T(z_0))^{n-1} / 2^{n-2}.$$

Further results in the weakly pseudoconvex finite-type case were obtained by Kohn. In [12] he gave a general method based on ideals of C^∞ -functions that are "subelliptic multipliers." Two special cases of this method are:

If $\Omega \subset \mathbf{C}^2$, and if $b\Omega$ is pseudoconvex and smooth and if $T(z_0) < \infty$, then (3) holds near z_0 . (4)

If $b\Omega$ is pseudoconvex and real-analytic and if $T(z_0) < \infty$, then (3) holds near z_0 . (5)

In the case of pseudoconvex boundaries, we can answer the question of when (3) holds for some $\varepsilon > 0$.

THEOREM [5]. *Assume that the boundary of Ω is pseudoconvex and smooth near z_0 . Then (3) holds for some $\varepsilon > 0$ if and only if $T(z_0) < \infty$.*

In [2], it was shown that $\varepsilon \leq 1/T(z_0)$. However examples show that if $\varepsilon(z_0) = \sup\{\varepsilon; (3) \text{ holds for } \varepsilon\}$, then it may happen that $\varepsilon < 1/T(z_0)$ [6] and that (3) sometimes may fail for $\varepsilon = \varepsilon(z_0)$ (based on unpublished examples).

The first part of the proof of the above theorem consists of reducing (3) down to the construction of a family of bounded plurisubharmonic functions with sufficiently large Hessian near the boundary. If g is a smooth function in \mathbf{C}^n , define the Hessian of g by

$$Hg(z; t) = \sum_{i,j} \frac{\partial^2 g}{\partial z_i \partial \bar{z}_j}(z) t_i \bar{t}_j.$$

If we use a plurisubharmonic function λ satisfying $|\lambda| \leq 1$ in Hörmander's weighted $\bar{\partial}$ -estimates [9], we obtain

$$\int_{b\Omega} H_r(z; U) dS + \int_{\Omega} H_\lambda(z; U) dV + \sum_{i,j=1}^n \left\| \frac{\partial U_i}{\partial \bar{z}_j} \right\|^2 \leq CQ(U, U). \quad (6)$$

Since $b\Omega$ is pseudoconvex and λ is plurisubharmonic, it follows that both $H_r(z; U)$ and $H_\lambda(z; U)$ are nonnegative. Hence, each of the three terms is dominated by $Q(U, U)$.

By standard techniques in partial differential equations, one may show that

$$\|U\|_{\varepsilon}^2 \leq C \left(\|U_b\|_{\varepsilon-1/2}^2 + \sum_{i,j} \left\| \frac{\partial U_i}{\partial \bar{z}_j} \right\| \right),$$

where U_b denotes the restriction of U to $b\Omega$. To estimate $\|U_b\|_{\varepsilon-1/2}^2$, we decompose U microlocally. Let $\varphi_k(t)$ satisfy $\sum_{k=1}^{\infty} \varphi_k^2(t) = 1$, where $\varphi_k(t) = 0$ if $t \leq 2^{k-1}$ ($k > 1$) or if $t \geq 2^{k+1}$, $k \geq 1$. Now define U_k by $\tilde{U}_k(\xi, r) = \tilde{\varphi}_k(|\xi|) \tilde{U}(\xi, r)$. It follows that

$$\|U_b\|_{\varepsilon-1/2}^2 \leq C \sum_k \|U_{k,b}\|^2 2^{2k(\varepsilon-1/2)}.$$

Let $S(\delta) = \{z; -\delta < r(z) < 0\}$. By averaging $\|U_{k,b}\|^2$ over $S(2^{-k})$, we see that

$$\|U_{k,b}\|^2 2^{2k(\varepsilon-1/2)} \leq \int_{S(2^{-k})} |U_k|^2 dV 2^{2k\varepsilon}.$$

Let us now make the following assumption:

ASSUMPTION. For all small $\delta > 0$ there exists a smooth plurisubharmonic function λ_{δ} on $\bar{\Omega}$ with $|\lambda_{\delta}| \leq 1$ and

$$H_{\lambda_{\delta}}(z; t) \geq c\delta^{-2\varepsilon}|t|^2, \quad z \in S_{\delta}. \quad (7)$$

If we apply (6) with $\lambda = \lambda_{\delta}$ and $\delta = 2^{-k}$, then we see that

$$2^{2k\varepsilon} \int |U_k|^2 dV \leq CQ(U_k, U_k).$$

Hence

$$\begin{aligned} \|U_b\|_{\varepsilon-1/2}^2 &\leq C \sum_k Q(U_k, U_k) \leq C \left(\sum_k \|\bar{\partial}\Phi_k U\|^2 + \|\vartheta\Phi_k U\|^2 \right) \\ &\leq C \left(Q(U, U) + \sum_k \|[\bar{\partial}, \Phi_k]U\|^2 + \|[\vartheta, \Phi_k]U\|^2 \right), \end{aligned}$$

where Φ_k denotes the multiplier operator associated with $\varphi_k(|\xi|)$. By standard techniques involving pseudodifferential operators, the sum of commutator terms can be bounded by $C\|U\|^2$. Thus the proof of (3) is reduced to constructing the family of functions λ_{δ} . We remark that in the proof we have essentially replaced the weight function $e^{-\lambda}$ in the Hörmander estimates by a pseudodifferential operator $e^{-\lambda(z, D)}$, where $\lambda(z, \xi) = \sum_{k=1}^{\infty} \lambda_{2^{-k}}(z) \varphi_k(|\xi|)$.

Thus we want to verify that the Assumption holds when $T(z_0) < \infty$. We begin with the simple case when $b\Omega$ is strongly pseudoconvex. Set $\lambda_{\delta}(z) = \exp(\delta^{-1}r(z))$. Then

$$H_{\lambda_{\delta}}(z; t) = \exp(\delta^{-1}r(z)) \left[\frac{1}{\delta} H_r(z; t) + \frac{1}{\delta^2} \left| \sum_{k=1}^n \frac{\partial r}{\partial z_k}(z) t_k \right|^2 \right].$$

Since strict pseudoconvexity means that $H_r(z; t) > 0$ when $\sum_k (\partial r / \partial z_k) t_k = 0$, it follows easily that the quantity in brackets is bounded below by $c\delta^{-1}|t|^2$. When $z \in S_{\delta}$, then $\exp(\delta^{-1}r(z)) \geq 1/e$. Thus the Assumption holds with $\varepsilon = 1/2$.

Let us consider the weakly pseudoconvex domain in \mathbb{C}^2 given by

$$\Omega = \{z; \operatorname{Re} z_1 + (\operatorname{Re} z_2)^{2m} < 0\}, \quad m > 1.$$

To verify the Assumption with $\varepsilon = 1/2m$, we must construct λ_δ satisfying

$$H_{\lambda_\delta}(z; t) \geq c\delta^{-1/m}, \quad z \in S_\delta. \quad (8)$$

If we set $\lambda_\delta = \exp(\delta^{-1}r(z))$, then (8) holds if $|\operatorname{Re} z_2| \geq \delta^{1/2m}$. When $|\operatorname{Re} z_2| \leq \delta^{1/2m}$, we must add a correction term. Let $\chi(t)$ satisfy $\chi(t) = 1$ if $t \geq 1$ and $\chi(t) = t$ if $t \leq 1/2$. Let

$$\lambda'_\delta(z) = \chi(|\delta^{-1/2m} \operatorname{Re} z_2|^2) + C' \exp(\delta^{-1}r(z)).$$

This correction term causes $H_{\lambda'_\delta}(z; t)$ to be of size $\delta^{-1/m}|t|^2$ when $|\operatorname{Re} z_2| \leq \frac{1}{2}\delta^{1/2m}$. But when $\frac{1}{2}\delta^{1/2m} \leq |\operatorname{Re} z_2| \leq \delta^{1/2m}$, the Hessian of the correction term is of order of magnitude $-C\delta^{-1/m}|t|^2$. If $z \in S_\delta$, then these terms can be dominated by the Hessian of $\exp(\delta^{-1}r(z))$. Hence (7) follows for λ'_δ . It still may be that λ'_δ is not plurisubharmonic when $r(z) < -\delta$. However, this can be corrected by setting $\lambda''_\delta(z) = \psi(\lambda'_\delta(z))$, where $\psi(t) \equiv 0$ if $t \leq \frac{1}{2}C'$ and $\psi''(t) > 0$ if $t > \frac{1}{2}C'$.

In order to generalize the above construction to the class of weakly pseudoconvex domains of finite type, we will define an invariant that decomposes the weakly pseudoconvex points into a finite number of sets, each of which can be treated locally in the same manner as above. Let Γ denote the set of n -tuples $\Lambda = (\lambda_1, \dots, \lambda_n)$ such that

$$1 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n \leq +\infty, \quad (9)$$

and for each k , either $\lambda_k = +\infty$, or there is a set of nonnegative integers a_1, \dots, a_k , with $a_k > 0$, such that $\sum_{j=1}^k a_j/\lambda_j = 1$. It is elementary to show that the set of possible entries λ_k which can occur in Λ is a discrete set accumulating only at $+\infty$. An element Λ in Γ is said to be distinguished if there exist holomorphic coordinates (z_1, \dots, z_n) with z_0 mapped to the origin such that

$$D^\alpha \bar{D}^\beta r(z_0) = 0 \quad \text{if} \quad \sum_{i=1}^n \frac{\alpha_i + \beta_i}{\lambda_i} < 1.$$

The multitype $M(z_0)$ is the smallest $M \in \Gamma$ (in the lexicographic sense) such that $M \geq \Lambda$ for every distinguished Λ . Recall that a submanifold M of a pseudoconvex boundary has holomorphic dimension zero if every nonzero holomorphic vector $L = \sum_j t_j \partial/\partial z_j$ that is tangent to M satisfies $\partial \bar{\partial} r(L, \bar{L}) > 0$. The invariant $M(z)$ has the following properties:

THEOREM [4]. *Let Ω be a pseudoconvex domain with smooth boundary. Then*

(i) *For each $z_0 \in b\Omega$, there is a neighborhood V such that for all $z \in V \cap b\Omega$, $M(z) \leq M(z_0)$ (in the lexicographic sense).*

(ii) *Suppose that $M(z_0) = (m_1, \dots, m_n)$ and that $m_n < \infty$. Then there exist a neighborhood U of z_0 and a submanifold M of $U \cap b\Omega$ of holomorphic dimension zero such that $\{z \in U \cap b\Omega; M(z) = M(z_0)\} \subset M$.*

(iii) If $z_0 \in b\Omega$ and $M(z_0) = (m_1, \dots, m_n)$, then there exist coordinates (z_1, \dots, z_n) about z_0 such that if $\sum(\alpha_i + \beta_i)/m_i < 1$, then $D^\alpha \bar{D}^\beta r(z_0) = 0$. If one of the entries in $M(z_0)$ equals $+\infty$, then these coordinates should be interpreted as formal power series.

(iv) If $M(z_0) = (m_1, \dots, m_n)$, then $m_i \leq T(z_0)$, $i = 1, \dots, n$.

Suppose now that $z_0 \in b\Omega$ satisfies $T(z_0) < \infty$. Then by combining D'Angelo's theorem with (9) and property (iv), it follows that there is a neighborhood V of z_0 such that for all $z \in V \cap b\Omega$, each entry m_i of $M(z) = (m_1, \dots, m_n)$ satisfies $m_i \leq T = (T(z_0))^{n-1}/2^{n-2}$. Since the set of possible entries is discrete, we conclude that $M(z)$ takes on only a finite set of values, say $M_1 < M_2 < \dots < M_N$. If S_k is defined by

$$S_k = \{z \in V \cap b\Omega; M(z) \geq M_k\},$$

then $V \cap b\Omega$ is stratified into the sets $S_k - S_{k+1}$, $k = 1, \dots, N$, each of which is locally contained in a manifold of holomorphic dimension zero. We remark that a similar stratification for real-analytic boundaries due to Diederich and Fornaess [7] was used by Kohn in [12] to prove (3).

Assume that we have already constructed λ_δ so that the Hessian of λ_δ is large off of S_k . Since $S_k - S_{k+1}$ is locally contained in a manifold M of holomorphic dimension zero, we may construct in a small tubular neighborhood of M a bounded function whose Hessian is large near M . If we add this function to λ_δ as in the example in \mathbb{C}^2 , then it follows that the Hessian of the new function λ'_δ is large off of S_{k+1} . Since $S_{N+1} = \emptyset$, we obtain by induction a family of functions with large Hessian near $b\Omega \cap V$. In order to prove the estimate (7), one must use the finite type assumption to do all of this in a quantitative form. The details become very technical, but the ideas are still the same.

REFERENCES

1. S. Bell and E. Ligocka, *A simplification and extension of Fefferman's theorem on biholomorphic mappings*, Invent. Math. **57** (1980), 283–289.
2. D. Catlin, *Necessary conditions for subellipticity of the $\bar{\partial}$ -Neumann problem*, Ann. of Math. (2) **117** (1983), 615–637.
3. —, *Global regularity of the $\bar{\partial}$ -Neumann problem*, Proc. Sympos. Pure Math., vol. 41, Amer. Math. Soc., Providence, R. I., 1984.
4. —, *Boundary invariants of pseudoconvex domains*, Ann. of Math. (2) **120** (1984), 529–586.
5. —, *Subelliptic estimates for the $\bar{\partial}$ -Neumann problem on pseudoconvex domains*, Preprint.
6. J. D'Angelo, *Real hypersurfaces, orders of contact, and applications*, Ann. of Math. (2) **115** (1982), 615–637.
7. K. Diederich and J. E. Fornaess, *Pseudoconvex domains with real-analytic boundary*, Ann. of Math. (2) **107** (1978), 371–384.
8. C. Fefferman, *The Bergman kernel and biholomorphic mappings of pseudoconvex domains*, Invent. Math. **26** (1974), 1–65.
9. L. Hörmander, *An introduction to complex analysis in several variables*, Van Nostrand, Princeton, N.J., 1966.

10. J. J. Kohn, *Harmonic integrals on strongly pseudoconvex manifolds*. I, Ann. of Math. (2) **78** (1963), 112–148; II, Ann. of Math. (2) **79** (1964), 450–472.
11. —, *Boundary behavior of $\bar{\partial}$ on weakly pseudoconvex manifolds of dimension two*, J. Differential Geom. **6** (1972), 523–542.
12. —, *Subellipticity of the $\bar{\partial}$ -Neumann problem on pseudoconvex domains: sufficient conditions*, Acta Math. **142** (1979), 79–122.
13. J. J. Kohn and L. Nirenberg, *Non-coercive boundary value problems*, Comm. Pure Appl. Math. **18** (1965), 443–492.

PURDUE UNIVERSITY, WEST LAFAYETTE, INDIANA 47907, USA

Extremal Functions in a Sharp Form of Sobolev Inequality

SUN-YUNG A. CHANG

Introduction. Let D be a domain in the n -dimensional space \mathbf{R}^n and $x \in \mathbf{R}^n$, $n \geq 2$. $\mathring{W}_q^1(D)$ denotes the Sobolev space obtained from the C^1 -functions $u(x)$ with compact support in D by completion with the norm

$$\|u\|_q = \left(\int_D |u_x|^q dx \right)^{1/q}.$$

The well-known Sobolev imbedding theorem states that for $1 < q < n$, $\mathring{W}_q^1(D)$ is contained boundedly in $L^p(D)$ ($\frac{1}{p} = \frac{1}{q} - \frac{1}{n}$). When $q = n$, Moser [22] sharpened an earlier result of Trudinger [28] and obtained

THEOREM 1. *Let $u \in \mathring{W}_n^1(D)$, $n \geq 2$, and $\int_D |u_x|^n dx \leq 1$. Then there exists a constant C_n which depends only on n such that*

$$\int_D e^{\alpha u^p} \leq C_n m(D) \quad (1)$$

where $p = n/(n-1)$, $\alpha \leq \alpha_n = n(\omega_{n-1})^{1/n-1}$, $m(D) = \int_D dx$, and ω_{n-1} is the $(n-1)$ -dimensional surface of the unit sphere. The integral on the left actually is finite for any such α , but if $\alpha > \alpha_n$ it can be made arbitrarily large by an appropriate choice of u .(*)

The remarkable fact is that the above inequality still holds for the critical value $\alpha = \alpha_n$. In this article, we will discuss the existence of an extremal function (when $D =$ unit ball in \mathbf{R}^n) for the Moser-Trudinger inequality, as well as related inequalities which occur in some problems in complex and real analysis, and indicate an application of a variant of the inequality (1) in a geometric problem raised by L. Nirenberg concerning Gaussian curvature on S^2 .

(*) A generalization of Theorem 1 to the Sobolev space $\mathring{W}_q^m(D)$ ($m < n$, $q = n/m$) has recently been done by D. Adams.

1. Extremal functions. If we take the domain $D = \mathbf{R}^n$, then for functions in $\mathring{W}_q^1(\mathbf{R}^n)$, $1 < q < n$, the Sobolev imbedding theorem says (for $\frac{1}{p} = \frac{1}{q} - \frac{1}{n}$)

$$\left(\int_{\mathbf{R}^n} |u|^p dx \right)^{1/p} \leq C(n, q) \left(\int_{\mathbf{R}^n} |u_x|^q dx \right)^{1/q}. \quad (2)$$

The best constant $C(n, q)$ and the extremal functions for the inequality (2) are known (cf. Hardy-Littlewood-Bliss [5], T. Aubin [1], and Talenti [30]). One extremal function is of the form

$$u(x) = (1 + \|x\|^{q/q-1})^{1-n/q} \quad (3)$$

and the extremal function for (2) is “unique” up to translation x_0 , and rescaling of $u(x)$:

$$u_{\lambda, x_0}(x) = (\lambda + \|x - x_0\|^{q/q-1})^{1-n/q}$$

for $\lambda \in \mathbf{R}$, $x_0 \in \mathbf{R}^n$. The precise form and the “uniqueness” of the extremal functions for this form of the Sobolev imbedding theorem has played an important role in establishing solutions for a number of nonlinear partial differential equations. (We refer the reader to the recent survey article of Brezis [6] for more complete references.) In particular when $q = 2$, $p = 2n/(n-2)$, and the Sobolev imbedding $\mathring{W}_2^1 \hookrightarrow L^{2n/(n-2)}$ fails to be compact; the function $u(x)$ in (3) was used critically in some nonlinear P.D.E. problems arising from differential geometry—for example, in Schoen’s work [26] of the Yamabe problem (cf. §3 of the article and, again, [6]).

The point we want to make here is that since the best constant $C(n, q)$ is unchanged through a dilatation argument when we restrict inequality (2) to functions $u \in \mathring{W}_q^1(\mathbf{B}^n)$ where \mathbf{B}^n is the unit ball in \mathbf{R}^n , by the uniqueness of the extremal functions (3) for the inequality (2) on \mathbf{R}^n , extremal functions for (2) *do not* exist for the Sobolev inequality (2) when we restrict to functions $u \in \mathring{W}_q^1(\mathbf{B}^n)$.

Thus it came as a surprise to us when, in [8], Carleson and I established the existence of an extremal function for Moser’s inequality (1) when $D = \mathbf{B}^n$ and $\alpha = \alpha_n$. Since the existence proof in [8] is quite indirect, we were not able to write down an explicit extremal function or to compute the best constant C_n in inequality (1). What we have actually shown is that if we let $K_n = 1 + e^{1+1/2+\dots+1/(n-1)}$, then K_n is the least number beyond which “compactness” can be established for sequences of functions in $\mathring{W}_n^1(\mathbf{B}^n)$, and $C_n > K_n$. This phenomenon is a particular example of the general phenomenon of compensated compactness which has recently appeared in a number of nonlinear problems [18, 6].

2. Some related inequalities. In this section we will mention inequalities similar to (1) which have occurred in some problems in real and complex analysis.

A. An inequality of A. Beurling. In the case $n = 2$, Moser’s inequality in Theorem 1 states that functions in $\mathring{W}_2^1(D)$ with finite Dirichlet integrals are

exponential square integrable. In Beurling's influential study [4] of majorization and the extremal length-area principle, he established the following estimate: If $f(z)$ is an analytic function on $|z| < 1$ and

$$\iint_{|z|<1} |f'(z)|^2 dx dy \leq \pi, \quad f(0) = 0, \quad (4)$$

then $\text{meas}\{\theta \mid |f(e^{i\theta})| \geq M\} \leq \pi e^{-M^2}$ for each $M > 0$, and the estimate is sharp.

In connection with Moser's inequality, it is natural to ask whether the function f satisfying (4) is $\exp |f(e^{i\theta})|^2$ integrable. (Beurling's weak-type estimate indicates that $\exp \alpha |f(e^{i\theta})|^2$ is integrable for any $\alpha < 1$, and the integral is not bounded over the whole class of functions satisfying (4) for all $\alpha > 1$.) That this is indeed the case can be verified either by a variation of Moser's proof of inequality (1) (in [9]) or by a complex analysis method based more on Beurling's original work on extremal length estimate (in [21]). We state this as

THEOREM 2. *There is a constant $C < \infty$ such that if f is analytic on the unit disk $|z| < 1$, $f(0) = 0$, and $\iint_{|z|<1} |f'(z)|^2 dx dy \leq \pi$, then*

$$\int_0^{2\pi} e^{|f(e^{i\theta})|^2} d\theta \leq C.$$

Theorem 2 has been generalized by M. Essén to cases where the area of the unit disc under the mapping function f (not counting multiplicity) is $\leq \pi$ (cf. [14]). He considers also problems where the integrand can be more general than $\exp(|f(e^{i\theta})|^2) d\theta$.

B. Functions with bounded square functions. Functions in the exponential square class have also appeared in some recent developments in classical harmonic analysis. To state the results, we need to introduce some notation. Given a locally integrable function f defined on \mathbf{R}^n , let $F(x, y)$ denote the harmonic extension of f to the upper half-space $\mathbf{R}_+^{n+1} = \{(x, y) : x \in \mathbf{R}^n, y > 0\}$. For each $0 < \gamma < \infty$, the classical Lusin area function is the square function defined as

$$S_\gamma f(x_0) = \left(\int_{\Gamma_\gamma(x_0)} |\nabla F(x, y)|^2 y^{1-n} dx dy \right)^{1/2},$$

and the nontangential maximal function f is defined as

$$N_\gamma f(x_0) = \sup_{(x, y) \in \Gamma_\gamma(x_0)} |F(x, y)|$$

where $\Gamma_\gamma(x_0) = \{(x, y) \in \mathbf{R}_+^{n+1}, |x - x_0| < \gamma y\}$. In the 1970s the fundamental results of Burkholder-Gundy-Silverstein ([7] for $n = 1$) and C. Fefferman-Stein ([15] for general n) state that for all $0 < p < \infty$, all γ_1, γ_2 , $\|N_{\gamma_1}(f)\|_p \sim \|S_{\gamma_2} f\|_p$, and thus, via the (generalized) Cauchy-Riemann equations, a function f with $S_\gamma f \in L^p$ is in the Hardy space $H^p(\mathbf{R}^n)$ (here, when $n > 1$, $H^p(\mathbf{R}^n)$ is the space defined by Stein-Weiss). This so-called real analysis definition of H^p was the starting point of some active research in H^p -theory in the past ten or so years. A natural question to ask then is: What happens when $p = \infty$? It is

relatively easy to see that the bounded function f always has its area function $S_\gamma f$ exponential square integrable (cf. [10, §1]) but may not be bounded. But for functions f with $S_\gamma f \in L^\infty$, a deep result of T. Wolff states that the best one can say is that f is (locally) exponentially square integrable. We state this as

THEOREM 3 [10, THEOREM 3.2]. *If $S_\gamma f \in L^\infty$, then there exists C_1, C_2 depending only on n and such that*

$$\sup_{Q: \text{cube}} \frac{1}{|Q|} \int_Q \exp \left(C_1 \frac{|f - f_Q|^2}{\|S_\gamma f\|_\infty^2} \right) < C_2$$

where $f_Q = (1/|Q|) \int_Q f$.

Wolff's original proof of Theorem 3 depends on the central limit theorem in probability. Later Rubin found an easier proof of the theorem in its dyadic martingale version. Theorem 3 also has been generalized to probabilistic square functions by R. Banuelos, to product domain (e.g., polydisc) square functions by J. Pipher [25], and to square functions corresponding to strictly elliptic operators in its divergence form by C. Sweezy [27].

3. A problem of prescribing Gaussian curvature on S^2 . On the standard two-sphere $S^2 = \{x \in \mathbf{R}^3 | x_1^2 + x_2^2 + x_3^2 = 1\}$ with metric $dS_0^2 = dx_1^2 + dx_2^2 + dx_3^2$, when the metric is subject to the conformal change $ds^2 = e^{2u} ds_0^2$, the Gaussian curvature of the new metric is determined by the equation

$$\Delta u + K e^{2u} = 1 \quad \text{on } S^2 \quad (5)$$

where Δ denotes the Laplacian relative to the standard metric. The question raised by L. Nirenberg is: Which functions K can be prescribed so that (5) has a solution u ? There is an obvious necessary condition implied by integration of (5) over the whole sphere (with $d\mu$ denoting the standard surface measure on S^2): $\int_{S^2} K e^{2u} d\mu = 4\pi$. Thus K must be positive somewhere. A further necessary condition has been noted in Kazdan-Warner [19]. For each eigenfunction x_j , with $\Delta x_j + 2x_j = 0$ ($j = 1, 2, 3$), the Kazdan-Warner condition states that

$$\int_{S^2} \langle \nabla K, \nabla x_j \rangle e^{2u} d\mu = 0, \quad j = 1, 2, 3.$$

Thus functions of the form $K = \psi \circ x_j$ (in particular, $K = 1 + \varepsilon x_j$ for any $\varepsilon > 0$), where ψ is any monotonic function defined on $[-1, 1]$, do not admit solutions. In [23], Moser approached this problem by the variational method, and he studied the functional

$$F[u] = \log \frac{1}{4\pi} \int_{S^2} K e^{2u} d\mu - \frac{1}{4\pi} \int_{S^2} |\nabla u|^2 d\mu - \frac{2}{4\pi} \int_{S^2} u d\mu. \quad (6)$$

He also proved that when K is an even function on S^2 (i.e., K is reflection symmetric about the origin), F achieves its maximum on even functions in $W_2^1(S^2)$, and hence its maximum u satisfies Euler's equation. Moser's proof in [23] was

based on a variant of his sharp inequality (1). On S^2 the precise inequality takes the following form:

THEOREM 1' [11]. *Given $u \in W_2^1(S^2)$, there is a universal constant C_0 such that*

$$\frac{1}{4\pi} \int_{S^2} \exp \left(\alpha(u - \bar{u})^2 / \int_{S^2} |\nabla u|^2 d\mu \right) d\mu \leq C_0 \quad (7)$$

for all $\alpha \leq 4\pi$ where $\bar{u} = (1/4\pi) \int_{S^2} u d\mu$.

Notice that as an immediate consequence of (7), we have a

COROLLARY. *For all $u \in W_2^1(S^2)$*

$$\frac{1}{4\pi} \int_{S^2} e^{2u} d\mu \leq C'_0 \exp \left(\frac{1}{\alpha} \int_{S^2} |\nabla u|^2 d\mu + \frac{2}{4\pi} \int_{S^2} u d\mu \right) \quad (8)$$

for all $\alpha \leq 4\pi$ for some constant $C'_0 \leq C_0$.^(**)

Hence it follows from (8) that for all $u \in W_2^1(S^2)$

$$F[u] \leq \log \max K + \log C'_0 + \left(\frac{1}{\alpha} - \frac{1}{4\pi} \right) \int_{S^2} |\nabla u|^2 d\mu.$$

Since $\alpha = 4\pi$ is allowed in (8), the functional $F[u]$, in particular, is bounded from above. The problem is this: Some maximizing sequence $\{u_k\}$ for the functional F may fail to be bounded in $W_2^1(S^2)$. (When $K \equiv 1$, one can examine to see that that is indeed the case.) But as pointed out in [23], for even functions u , the best exponent in the inequality is $\alpha = 8\pi$. Thus the functional F achieves its maximal over the even functions when K is even. To better explain the doubling of the best exponent α when u is even, we observed the following relationship between α and the isoperimetric inequality [11].

LEMMA. *Suppose D is a bounded, piecewise C^1 domain in the plane with finite vertices. Let θ_D be the minimum interior angle at the vertices of D . For each $u \in C^2(\bar{D})$, and each real number M , denote $L_M =$ arc length of the set $\{u = M\}$, $A_M =$ area of the set $\{u \geq M\}$, and let $\alpha(u, D) = \lim_{A_M \rightarrow 0} L_M^2 / A_M$. Then $\alpha(u, D) \geq 2\theta_D$ and there exists a constant C_D such that for all u we have $(\bar{u} = (1/m(D)) \int_D u)$*

$$\left(\int_D \exp(\alpha(u, D)(u - \bar{u})^2) / \int_D |\nabla u|^2 \right) \leq C_D. \quad (9)$$

This lemma could be reformulated for domains in the sphere S^2 , also. There one can easily check for general $u \in C^1(S^2)$, $\alpha(u, S^2) \geq 4\pi$, while for even u , $\alpha(u, S^2) \geq 8\pi$. Notice also, when D is a smooth domain in S^2 , $\theta_D = \pi$ and the best exponent in (9) is 2π . Following the same approach of Moser by taking the

^(**)In connection with string theory in physics, Onofri [24] proved that the best constant C'_0 in (8) is 1.

supremum over the corresponding functional F , we get the following solution for equation (5) for the corresponding Neumann problem:

THEOREM 4. *Suppose K is a smooth function on a smooth domain D in S^2 ; for the equation*

$$\begin{aligned}\Delta u + Ke^{2u} &= 1 & \text{on } D, \\ \partial u / \partial n &= 0 & \text{on } \partial D\end{aligned}\tag{5}'$$

(where $\partial/\partial n$ denotes metric unit normal derivative) to have a solution, it suffices that

- (a) when $|D| < 2\pi$, K be positive somewhere on D ;
- (b) when $D =$ a hemisphere H , K satisfies

$$\frac{1}{2\pi} \int_H K d\mu > \max_{x \in \partial H} (K(x), 0).\tag{10}$$

Notice that case (b) also gives a sufficient condition for the original Nirenberg problem for those K satisfying a reflection symmetry about some plane (e.g., $K(x_1, x_2, x_3) = K(x_1, x_2, -x_3)$). In the case where K possesses rotational symmetry (around some axis), some sufficient conditions similar to (10) were given in Hong [17].

For all the sufficient conditions in [23, 11, 17], which we presented here, solutions obtained for equation (5) were global maxima of the functional F restricted to some suitable subspace of W_2^1 ; but it is actually the case that when K is a positive function, no solution to (5) can be a local maximum (i.e., index zero solution) unless K is identically a constant. This follows from a second variation computation coupled with an eigenvalue estimate of Hersch [16]. Thus, to obtain a general solution of the equation (5), we need to capture saddle points (i.e., unstable critical points) of the functional F ; i.e., we need to look for some max-min schemes for F . Such attempts are made in [12], and we have

THEOREM 5. *Suppose K is a smooth positive function with two nondegenerate local maximum (which we may assume w.l.o.g.) located at the north and south poles N, S . Let ϕ_t be the one-parameter group of conformal transformations given in terms of stereographic complex coordinates (with $z = \infty$ corresponding to N and $z = 0$ corresponding to S) by $\phi_t(z) = tz$, $0 < t < \infty$. Assume*

$$\inf_{0 < t < \infty} \frac{1}{4\pi} \int_{S^2} K \circ \phi_t d\mu > \max_{\substack{\nabla K(Q)=0 \\ Q \neq N, S}} K(Q).\tag{11}$$

Then (5) admits a solution.

REMARK. (11) is an analytic condition about the distribution of K which can be verified, for example, if K has nondegenerate local maximum points at N, S , as well as the following properties: (1) K has (suitably) small variation in the region $\{|x_3| > \varepsilon\}$. (2) All other critical points of K occur in the strip $\{|x_3| < \varepsilon\}$, and there the critical values of K are significantly lower than the minimum value of K on $\{|x_3| > \varepsilon\}$.

THEOREM 6. *Let K be a positive smooth function with only nondegenerate critical points, and, in addition, $\Delta K(Q) \neq 0$ where Q is any critical point. Suppose there are $p + 1$ local maximum points of K , and q saddle points of K with $\Delta K(Q) < 0$. If $q \neq p$, then K admits a solution to equation (5).*

As explained above, we need to look for saddle points of F . Thus the main idea in the proofs of Theorems 5 and 6 is to analyze when a maximizing max-min sequence fails to be compact (as in common in this type of nonlinear P.D.E. problem). For a function $u \in W_2^1(S^2)$, define

$$S[u] = \frac{1}{4\pi} \int_{S^2} |\nabla u|^2 d\mu + \frac{2}{4\pi} \int_{S^2} u d\mu.$$

It turns out that for a function $u \in W_2^1(S^2)$ normalized by $(1/4\pi) \int_{S^2} e^{2u} d\mu = 1$, $S[u]$ measures, in a very precise way, how “close” u can be compared to the standard solution of $\Delta u + e^{2u} = 1$ (i.e., $u = \frac{1}{2} \log |d\phi|$ for some conformal transformation of S^2 ; for such functions u , $S[u] = 0$). Since any maximizing sequences u_j for a max-min problem will automatically satisfy the condition $S[u_j] \leq C$ (for some constant C , after normalizing the sequence by $(1/4\pi) \int_{S^2} e^{2u_j} d\mu = 1$), if we do not have convergence, then we have good control of the behavior of concentration of $\{e^{2u_j}\}$; the hypothesis placed on K in Theorems 5 and 6 are used to rule out this type of behavior for maximizing sequences of some max-min schemes.

While all the results cited above give sufficient conditions for the existence of a solution of equation (5), there is another result of Kazdan-Warner [19] which states that for any K positive somewhere on S^2 , there always exists some diffeomorphism ϕ such that the equation $\Delta u + K \circ \phi e^{2u} = 1$ is solvable. It is therefore of interest to find some analytic conditions on the class of functions K which is topologically simple (e.g., K has only a global maximum and a global minimum) that ensures existence of a solution of (5).

In related developments, there is an analogous equation for prescribing scalar curvature on a compact manifold M of dimension n , $n \geq 3$. The corresponding equation for (5) becomes

$$\Delta u + R u^{(n+2)/(n-2)} = \frac{n-2}{4(n-1)} R_0 u \quad (12)$$

where R_0 is the scalar curvature of the underlying metric dS_0^2 and R is the prescribed scalar curvature of the conformally related metric $dS^2 = u^{4/(n-2)} dS_0^2$. As we have mentioned in §1 of this article, the difficulty in solving (12) for some positive solution u again lies in the failure of the compactness of the Sobolev imbedding $\overset{\circ}{W}_2^1 \hookrightarrow L^{2n/(n-2)}$. When $R = \text{constant}$, this was the Yamabe problem and was solved in partial cases by Yamabe [31], Trudinger [29], T. Aubin [2], and in the remaining cases, more recently by Schoen [26]. For the analogous problem of prescribing R on S^n ($n \geq 3$) with dS_0^2 the standard metric on S^n , Escobar and Schoen [13] gave the analogue of Moser’s theorem (for even functions on S^2); Bahri and Coron [3] have announced an analogue of Theorem 6 on S^3 .

REFERENCES

1. T. Aubin, *Problèmes isopérimétriques et espaces de Sobolev*, J. Differential Geom. **11** (1976), 573–598.
2. —, *Equations différentielles non linéaires et problèmes de Yamabe concernant la courbure scalaire*, J. Math. Pures Appl. **55** (1976), 269–296.
3. A. Bahri and J. M. Coron, *Une théorie des points critiques à l'infini pour l'équation de Yamabe et le problème de Kazdan-Warner*, C. R. Acad. Sci. Paris Sér. I Math. **15** (1985), 513–516.
4. A. Beurling, *Études sur un problème de majoration*, Thèse, Almqvist and Wiksell, Uppsala, 1933.
5. G. Bliss, *An integral inequality*, J. London Math. Soc. **5** (1930), 40–46.
6. H. Brezis, *Some variational problems with lack of compactness*, Proc. Berkeley Sympos. on Nonlinear Functional Analysis, 1985.
7. D. L. Burkholder, R. F. Gundy, and M. L. Silverstein, *A maximal function characterization of the class H^p* , Trans. Amer. Math. Soc. **157** (1971), 137–153.
8. L. Carleson and S. Y. A. Chang, *On the existence of an extremal function for an inequality of J. Moser*, Bull. Sci. Math. (2) (to appear).
9. S. Y. A. Chang and D. E. Marshall, *On a sharp inequality concerning the Dirichlet integral*, Amer. J. Math. **107** (1985), no. 5, 1015–1033.
10. S. Y. A. Chang, T. Wolff, and M. Wilson, *Some weighted norm inequalities concerning the Schrödinger operators*, Comment. Math. Helv. **60** (1985), 217–246.
11. S. Y. A. Chang and P. Yang, *Conformal deformation of metrics on S^2* , Preprint.
12. —, *On prescribing Gaussian curvature on S^2* , Preprint.
13. J. F. Escobar and R. Schoen, *Conformal metrics with prescribed scalar curvature*, Preprint.
14. M. Essén, *Sharp estimates of uniform harmonic majorants in the plane*, Ark. Mat. (to appear).
15. C. Fefferman and E. Stein, *H^p spaces of several variables*, Acta Math. **129** (1972), 137–192.
16. J. Hersch, *Quatre propriétés isopérimétriques de membranes sphériques homogène*, C. R. Acad. Sci. Paris **270** (1970), A1645–A1648.
17. Chongwei Hong, *A best constant and the Gaussian curvature*, Preprint.
18. S. Jacobs, *An isoperimetric inequality for functions analytic in multiply connected domains*, Mittag Leffler Report, 1972.
19. J. Kazdan and F. Warner, *Curvature functions for compact 2-manifold*, Ann. of Math. (2) **99** (1974), 14–47.
20. —, *Existence and conformal deformation of metrics with prescribed Gaussian and scalar curvature*, Ann. of Math. (2) **101** (1975), 317–331.
21. D. E. Marshall, *A new proof of a sharp inequality concerning the Dirichlet integral*, Ark. Mat. (to appear).
22. J. Moser, *A sharp form of an inequality by N. Trudinger*, Indiana Univ. Math. J. **20** (1971), no. 11, 1077–1091.
23. —, *On a non-linear problem in differential geometry, dynamical systems*, M. Peixoto, Editor, Academic Press, New York, 1973.
24. E. Onofri, *On the positivity of the effective action in a theory of random surface*, Comm. Math. Phys. **86** (1982), 321–326.
25. J. Pipher, Thesis, U.C.L.A., 1985.
26. R. Schoen, *Conformal deformation of a Riemannian metric to constant scalar curvature*, J. Differential Geom. **20** (1985), 479–495.
27. C. Sweezy, *L-Harmonic functions and the exponential square class*, Thesis, U.C.L.A., 1986.
28. N. S. Trudinger, *On imbedding into Orlicz spaces and some applications*, J. Math. Mech. **17** (1967), 473–484.

- 29. —, *Remarks concerning the conformal deformation of Riemannian structures on compact manifolds*, Ann. Scuola Norm. Sup. Pisa **22** (1968), 265–274.
- 30. G. Talenti, *Best constant in Sobolev inequality*, Ann. Mat. Pura Appl. **110** (1976), 353–372.
- 31. H. Yamabe, *On the deformation of Riemannian structures on compact manifolds*, Osaka J. Math. **12** (1960), 21–37.

UNIVERSITY OF CALIFORNIA, LOS ANGELES, CALIFORNIA 90024, USA

Chirurgie sur les applications holomorphes

A. DOUADY

Introduction. Si f est une fraction rationnelle complexe, on note $J(f)$ son *ensemble de Julia*: l'application f opère sur la sphère de Riemann $\overline{\mathbf{C}}$, et $J(f)$ est l'ensemble des z tels que les itérées f^n de f ne forment une famille équicontinue sur aucun voisinage de z . Si f est un polynôme, $J(f)$ est la frontière de l'ensemble $K(f)$ des z tels que la suite $f^n(z)$ reste bornée (*ensemble de Julia rempli*).

Soit f un polynôme monique de degré d tel que $K(f)$ soit connexe. Sur $\mathbf{C} - K(f)$, l'application f est holomorphiquement conjuguée à $f_0: z \rightarrow z^d$ sur $\mathbf{C} - \overline{D}$. Si en outre $K(f)$ est localement connexe, la représentation conforme $\psi: \mathbf{C} - \overline{D} \rightarrow \mathbf{C} - K(f)$ se prolonge au bord, et définit une surjection $\gamma_f: \mathbf{T} = \mathbf{R}/\mathbf{Z} \rightarrow \partial K(f)$ (*lacet de Caratheodory*). On a $f(\gamma_f(t)) = \gamma_f(d \cdot t)$.

On note P_c le polynôme $z \rightarrow z^2 + c$, et on écrit K_c pour $K(P_c)$. L'ensemble des c pour lesquels K_c est connexe est noté M (*ensemble de Mandelbrot*). Les figures 1, 2 et 3 représentent K_c pour $c_1 = -0,122561 + 0,744861i$, $c_2 = -0,156520 + 1,032247i$, $c_3 = -0,153941 + 1,03770i$ respectivement. Ce sont des valeurs de c pour lesquelles le point critique 0 est périodique de période $k_1 = 3$, $k_2 = 4$ et $k_3 = 12$ pour P_c . L'ensemble K_{c_1} est connu comme le "*lapin*."

Pour c tel que 0 soit périodique de période k pour P_c , notons $U_n(c)$ la composante connexe de l'intérieur de K_c contenant $P_c^n(0)$. On a $P_c^k(U_n) = U_{n+k} = U_n$. Sur un voisinage de la fermeture de U_n , l'application P_c^k est topologiquement conjuguée à $z \rightarrow z^2$ au voisinage de \overline{D} . Toute composante de l'intérieur de K_c tombe en un temps fini sur le cycle des $U_n(c)$.

On peut décrire le système dynamique défini par P_{c_3} opérant sur \mathbf{C} en disant qu'il est obtenu à partir de celui de P_{c_2} en remplaçant la fermeture de chaque composante de l'intérieur de K_{c_2} par un lapin. Sur le lapin contenant 0, l'application $P_{c_3}^4$ est topologiquement conjuguée à P_{c_1} sur K_{c_1} (conjugaison holomorphe à l'intérieur). On exprime cette situation en disant que P_{c_3} est obtenu en *modulant* (tuning) P_{c_2} par P_{c_1} . Dans M , il y a une copie homéomorphe M_{c_2} de M centrée en c_2 , et c_3 est le point de M_{c_2} correspondant à c_1 .

Dans cet exposé, nous allons essayer de faire un inventaire des cas découverts récemment ou l'on peut ainsi enlever à un système dynamique holomorphe un

morceau, le remplacer par un morceau d'un autre système dynamique holomorphe, et obtenir un nouveau système dynamique, holomorphe malgré la couture. Cette idée remonte au théorème d'uniformisation simultanée de L. Bers, et au théorème de recombinaison de Klein-Maskit, mais dans ces deux cas le temps est remplacé par un groupe dénombrable, et nous n'en parlerons pas ici. Tous les résultats obtenus utilisent le Théorème d'intégrabilité d'Ahlfors-Bers (Measurable Riemann-Mapping Theorem). C'est Sullivan qui a inauguré l'usage de ce théorème pour les systèmes dynamiques holomorphes à temps discret, c'est à dire pour l'itération des fractions rationnelles. Tous les résultats obtenus par l'auteur l'ont été en collaboration avec J. H. Hubbard.

Nous allons examiner successivement les applications à allure polynomiale et les opérations de modulation et d'accouplement. Puis nous décrirons un travail de M. Shishikura, qui a affiné la technique, et enfin deux applications, par Bodil Branner et par l'auteur, qui montrent la souplesse à laquelle on parvient.

Applications à allure polynomiale. Une *application à allure polynomiale* (polynomial-like mapping) est une application holomorphe propre $f: U' \rightarrow U$, où U et U' sont des ouverts de \mathbb{C} isomorphes à D , avec U' relativement compact dans U . On note alors $K(f)$ l'ensemble des z tels que $f^n(z)$ soit défini et appartienne à U' pour tout n .

Soient $f: U' \rightarrow U$ et $g: V' \rightarrow V$ deux applications à allure polynomiale, avec $K(f)$ et $K(g)$ connexes. Une *équivalence holomorphe* (resp. *quasi-conforme*) entre f et g est un isomorphisme analytique (resp. un homéomorphisme quasi-conforme) $\phi: U_1 \rightarrow V_1$, où U_1 et V_1 sont des voisinages de $K(f)$ et $K(g)$, tel que $g \circ \phi = \phi \circ f$ sur $U'_1 = f^{-1}(U_1)$. Une *équivalence hybride* est une équivalence quasi-conforme ϕ telle que $\partial\phi = 0$ presque partout sur $K(f)$. Une *équivalence extérieure* est un isomorphisme analytique $\phi: U_1 - K(f) \rightarrow V_1 - K(g)$ tel que $g \circ \phi = \phi \circ f$ sur $U'_1 - K(f)$.

THÉORÈME. *Etant donné deux applications à allure polynomiale $f: U' \rightarrow U$ et $g: V' \rightarrow V$ de même degré d , avec $K(f)$ et $K(g)$ connexes, il existe une application à allure polynomiale $h: W' \rightarrow W$, hybridement équivalente à f et extérieurement équivalente à g .*

Autrement dit, on peut recoller les systèmes dynamiques $K(f)$ muni de f et $V - K(g)$ muni de g en un système dynamique holomorphe.

PRINCIPE DE LA DEMONSTRATION. Quitte à rétrécir, on peut supposer que U et V sont à bord \mathbb{R} -analytique, ainsi que U' et V' , et que f et g se prolongent au bord. Soit ϕ_0 un difféomorphisme de ∂U sur ∂V . On peut trouver un difféomorphisme ϕ_1 de $\partial U'$ sur $\partial V'$ tel que $g \circ \phi_1 = \phi_0 \circ f$, puis étendre (ϕ_1, ϕ_0) en un difféomorphisme ϕ de $\bar{U} - U'$ sur $\bar{V} - V'$. On note σ la structure complexe $\phi^* \sigma_0$ sur $\bar{U} - U'$, où σ_0 est la structure standard. On peut étendre σ en une structure presque complexe sur $U - K(f)$ invariante par f , puis à U tout entier en prenant σ_0 sur $K(f)$. On note encore σ la structure obtenue.

D'après le Théorème d'Ahlfors-Bers, on peut trouver un homéomorphisme quasi-conforme ψ de U sur un ouvert W de \mathbb{C} tel que $\psi^*\sigma_0 = \sigma$. L'application $h = \psi \circ f \circ \psi^{-1}: W' \rightarrow W$, définie sur $W' = \psi(U')$, répond à la question.

REMARQUE. En degré 2, l'application h ainsi obtenue est unique à équivalence holomorphe près. En degré supérieur, le choix de la classe de ϕ , une fois choisis ϕ_0 et ϕ_1 , produit $d - 1$ classes d'équivalence holomorphe pour h .

COROLLAIRE. (*Théorème de redressement, straightening theorem.*) *Toute application à allure polynomiale de degré d admet une équivalence hybride avec un polynôme de degré d .*

PRINCIPE DE LA DEMONSTRATION. Parmi les applications à allure polynomiale de degré d , celles qui sont holomorphiquement équivalentes à des polynômes sont celles qui sont extérieurement équivalentes à $z \rightarrow z^d$.

REMARQUES. (1) Avec les définitions ci dessus, ceci ne s'applique qu'aux cas où $K(f)$ est connexe, mais on peut les modifier de façon à couvrir le cas général.

(2) Si on veut un polynôme monique centré, on a $d - 1$ choix si K est connexe, une infinité sinon.

(3) Soit $\mathbf{f} = (f_\lambda)$ une famille d'applications à allure polynomiale de degré d dépendant analytiquement d'un paramètre λ . Si $d = 2$, on peut choisir pour chaque λ un $c = \chi(\lambda)$ tel que P_c rectifie f_λ , de façon que χ soit continue (et même quasi-régulière sous certaines hypothèses, qui sont probablement inutiles en fait).

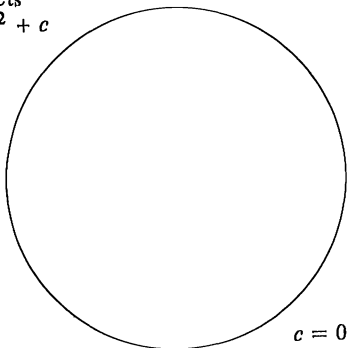
L'énoncé analogue en degré ≥ 3 est faux, même si on se restreint à l'ensemble $M_{\mathbf{f}}$ des valeurs de λ pour lesquelles $K(f_\lambda)$ est connexe. Ceci n'est pas dû aux $d - 1$ choix possibles: l'ensemble $K(f_\lambda)$ dépend de façon discontinue de λ en chaque point où f_λ a un cycle parabolique, et il en est de même de σ_λ et ψ_λ . Ce n'est que par miracle que, en degré 2, on peut récupérer la continuité pour $P_\lambda = \psi_\lambda \circ f_\lambda \circ \psi_\lambda^{-1}$.

(4) En degré 2, si λ parcourt un disque Λ dans lequel $M_{\mathbf{f}}$ est compact, cet ensemble $M_{\mathbf{f}}$, muni de χ , est un revêtement ramifié de M . Le degré δ de ce revêtement ramifié (*degré paramétrique*) est égal au nombre de tours que fait la valeur critique autour du point critique quand λ longe le bord de Λ . Si $\delta = 1$, l'application χ est un homéomorphisme, et on dit que \mathbf{f} est une famille *Mandelbrotesque*.

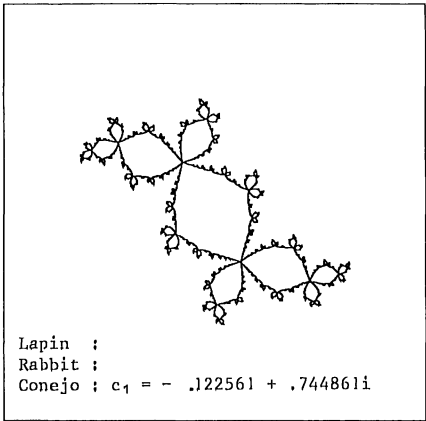
Modulation en degré 2. En degré 2, nous avons le résultat suivant: Soit c_0 tel que 0 soit périodique de période k pour P_{c_0} . Notons W la composante de l'intérieur de M qui contient c_0 , et c_1 la *racine* de cette composante (point du bord de W où le multiplicateur du cycle attractif devient égal à 1). On dit que W est une composante *primitive* si le bord de W a en c_1 un point de rebroussement; sinon W est rattachée en c_1 à une composante de période plus petite.

THÉORÈME DE MODULATION. (a) *Si W est une composante primitive, on peut trouver un voisinage Λ de \overline{W} , et deux familles d'ouverts $(U_c, U'_c)_{c \in \Lambda}$ telles que les $f_c = P_c^k: U'_c \rightarrow U_c$ forment une famille Mandelbrotesque.*

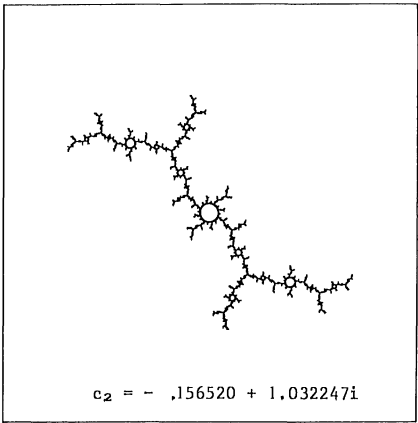
Julia sets
 $z \mapsto z^2 + c$



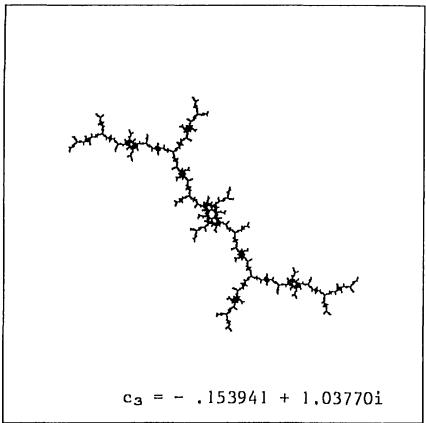
$c = 0$



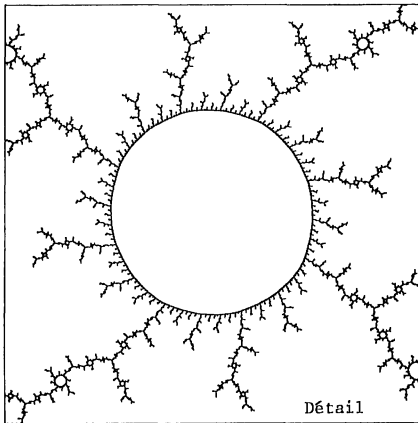
Lapin :
Rabbit :
Conejo : $c_1 = -.122561 + .744861i$



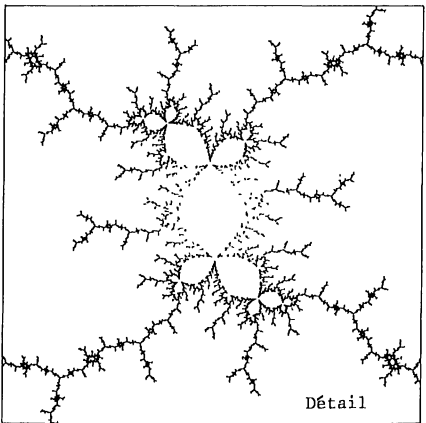
$c_2 = -.156520 + 1.032247i$



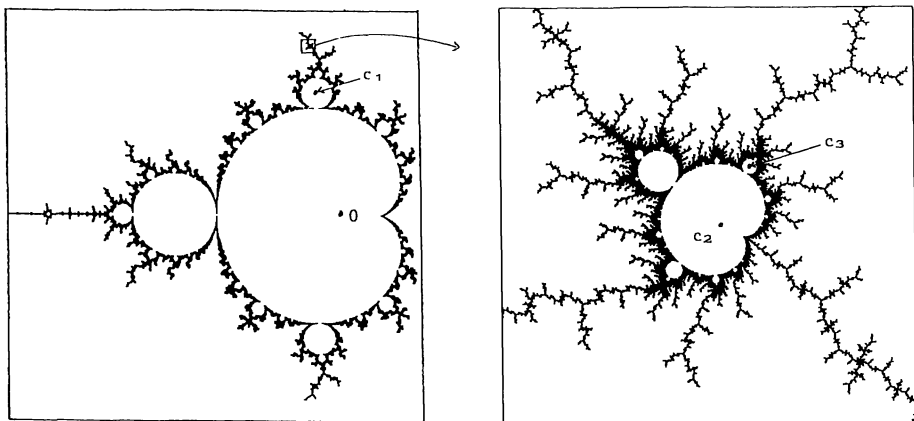
$c_3 = -.153941 + 1.03770i$



Détail



Détail



(b) Si W n'est pas primitive, on peut seulement construire Λ , let U_c et les U'_c de façon que Λ soit un voisinage de $\overline{W} - c_1$, et que χ induise un homéomorphisme de M_f sur $M - \{1/4\}$.

Pour $x \in M$, le point $\chi^{-1}(x)$ de Λ est appelé le *modulé* de c_0 par x (c_0 tuned by x). Dans le cas non primitif, l'application χ^{-1} admet un prolongement continu appliquant $\frac{1}{4}$ sur c_1 , ce qui permet de définir le modulé pour tout x de M .

La démonstration consiste à construire Λ , les U_c et les U'_c . Cette construction utilise les propriétés combinatoires des rayons externes décrites dans [6]. Comme elle n'est pas publiée, nous en donnons ici un plan assez détaillé, en nous basant sur des notes de P. Lavaurs. Donnons d'abord quelques définitions:

Soit K un compact de \mathbb{C} . Il existe un unique couple (r, ϕ) tel que ϕ soit un isomorphisme de $\mathbb{C} - K$ sur $\mathbb{C} - \overline{D}_r$ tangent à l'identité à l'infini. On dit que $r(K) = r$ est le *rayon de capacité* (ou *diamètre transfini*) de K , les *équipotentiellles* et les *rayons externes* de K sont les images réciproques des cercles centrés en 0 et des demi-droites. Le *potentiel* créé par K est $\text{Log}|\phi|$. L'*argument* d'un rayon externe est son argument à l'infini, i.e., l'argument de la demi-droite correspondante. Les arguments sont comptés en prenant comme unité le tour complet et non le radian. Si le rayon externe $R(K, t)$ d'argument t aboutit en un point x de K , on dit que x admet t comme argument externe; si K est localement connexe cela équivaut à $\gamma_K(t) = x$, ou γ_K est le lacet de Caratheodory.

On sait que M est un compact connexe de rayon de capacité 1. On ne sait pas démontrer qu'il est localement connexe, mais on sait que c_1 admet dans M deux arguments externes de la forme $\theta_s = p_s(2^k - 1)$, $s = 1, 2$ (k est la période de 0 pour P_{c_0}). On peut supposer $0 < \theta_1 < \theta_2 < 1$ (en excluant le cas trivial $c_0 = 0$).

Pour $c \in \mathbb{C}$, on peut définir le rayon externe $R(K_c, t)$ même si c n'est pas dans M , sauf si l'argument externe de c relativement à M est de la forme $2^n \cdot t$. En effet, on peut définir ϕ_c au voisinage de l'infini comme la fonction conjuguant P_c à P_0 , le potentiel défini au voisinage de l'infini par $G = \text{Log}|\phi|$ se prolonge à $\mathbb{C} - K_c$ grâce à l'équation fonctionnelle $G(z) = \frac{1}{2}G(P_c(z))$, et on peut prolonger

$R(K_c, t)$ comme ligne orthogonale aux équipotentielles, sauf s'il bute sur un col de potentiel.

Fixons $c = c_0$, écrivons K pour K_{c_0} , etc. Les rayons externes $R(K, \theta_1)$ et $R(K, \theta_2)$ aboutissent en un même point y_1 de ∂U_1 . Notons A l'ouvert limité par $L = R(K, \theta_1) \cup R(K, \theta_2) \cup \{y_1\}$ et une équipotentielle de niveau η . Notons A' la composante connexe de $P^{-k}(A)$ qui contient U_1 .

LEMME 1. *L'ouvert A' est contenu dans A . Il est limité par quatre rayons externes, deux d'arguments θ_1, θ_2 et deux autres d'arguments θ'_1 et θ'_2 , et deux arcs de l'équipotentielle de niveau $\eta' = 1/2^k \cdot \eta$. L'application P^k induit une application propre de degré 2 de A' dans A .*

Notons Λ_1 l'ouvert de \mathbb{C} limité par les rayons externes d'argument θ_1, θ_2 de M , et une équipotentielle de M de niveau $\eta' < \eta$, et contenant W .

LEMME 2. *Pour $c \in \Lambda$, les rayons externes $R(K_c, \theta_1)$ et $R(K_c, \theta_2)$ aboutissent en un même point $y_1(c)$; les rayons $R(K_c, \theta'_1)$ et $R(K_c, \theta'_2)$ aboutissent en un même point y'_1 . Ceci permet de définir des ouverts $A(c)$ et $A'(c)$. L'ouvert $A'(c)$ est contenu dans $A(c)$, et P_c^k induit une application propre de degré 2 de $A'(c)$ dans $A(c)$.*

L'application $P_c^k: A'(c) \rightarrow A(c)$ n'est pas à allure polynomiale, car $A'(c)$ n'est pas relativement compact dans $A(c)$. En effet, ils ont en commun dans leur frontière le point $y_1(c)$ et des arcs de $R(K_c, \theta_1)$ et $R(K_c, \theta_2)$. On remédie à cela en deux temps. Dans le premier, on remplace les rayons externes par des "paraboles de potentiel" d'équation $h = \varepsilon(t - \theta_s)^2$, où t est l'argument externe et h le potentiel, la constante ε choisie petite et du signe qu'il faut pour que l'ouvert modifié $A_1(c)$ soit contenu dans $A(c)$. On pose $A'_1(c) = A'(c) \cap P_c^{-k}(A_1(c))$. L'ouvert $A'_1(c)$ est encore contenu dans $A_1(c)$, et ces deux ouverts n'ont plus en commun dans leur frontière que le point $y_1(c)$. Jusqu'ici, la construction peut se faire pour c dans un ouvert Λ_2 obtenu en modifiant Λ de la même façon, et également pour $c = c_1$. Pour le deuxième temps, on définit U_c en ajoutant à $A_1(c)$ un disque $\Delta(c)$ centré en $y_1(c)$ et en posant $U_c = A'_1(c) \cup \Delta'(c)$, où $\Delta'(c)$ est la composante connexe de $P_c^{-k}(\Delta(c))$ contenant $y_1(c)$. Pour $c \in \Lambda_2$, si on a choisi $\Delta(c)$ assez petit, l'application P_c^k induit une application à allure polynomiale $f_c: U'_c \rightarrow U_c$.

Si W est une composante primitive, on peut effectuer ce deuxième temps aussi pour $c = c_1$, car au voisinage de y_1 le complémentaire de $A_1(c)$ est contenu dans l'unique interpétale de la fleur de Fatou, où P_c^k a un effet faiblement répulsif. Par stabilité, on obtient finalement des ouverts U_c, U'_c définissant une famille \mathbf{f} d'applications à allure polynomiale induites par les P_c^k , pour c parcourant un ouvert Λ réunion de Λ_2 et d'un petit voisinage de c_1 . On calcule le degré paramétrique et on montre que cette famille est Mandelbrotesque.

Si W n'est pas primitive, on a seulement une famille $\mathbf{f} = (f_c)$ indexée par $\Lambda = \Lambda_2$. Il faut alors montrer que $\chi(c)$ tend vers $\frac{1}{4}$ quand c tend vers c_1 dans M_f . Ceci peut se faire en utilisant une majoration de Yoccoz de la "taille des

membres de M ." On voit alors que M_f est un revêtement ramifié de degré fini de $M - \{\frac{1}{4}\}$, et il ne reste plus qu'à montrer que ce degré est 1.

Modulation en degré supérieur. Soit P un polynôme de degré d admettant un cycle superattractif ζ d'ordre k . On suppose que ζ contient un seul point critique α et on note δ le degré local de P en α (c'est à dire que la multiplicité de α comme zéro de P' est $\delta - 1$). On suppose également que tous les points critiques de P sont prépériodiques, et qu'aucun autre que α ne tombe sur le cycle ζ . La fermeture du bassin immédiat U_α de α est alors homéomorphe au disque fermé, et sur un voisinage de cette fermeture, P^k induit une application à allure polynomiale hybridement équivalente à $z \rightarrow z^\delta$. Sur ∂U_α , il y a $\delta - 1$ points fixes de P^k . Choisissons l'un d'eux comme point de base.

Soit Q un polynôme monique de degré δ tel que $K(Q)$ soit connexe et localement connexe. On peut définir un système dynamique en remplaçant la fermeture de chaque composante de l'intérieur de $K(P)$ qui tombe sur U_α par une copie de $K(Q)$, en prenant soin de recoudre le point de base au point de $K(Q)$ d'argument externe 0. Si ce système dynamique est topologiquement équivalent à celui défini par un polynôme P_Q , avec conjugaison holomorphe la ou c'est possible, on dit que P_Q est un *modulé* de P par Q .

Si tous les points critiques de P et Q sont périodiques (ou prépériodiques mais tombant sur un cycle superattractif) on peut démontrer l'existence du modulé en utilisant le Théorème de Thurston caractérisant topologiquement les fractions rationnelles parmi les revêtements ramifiés à ensemble postcritique fini. Un résultat démontré indépendamment par Israel Bernstein et Silvio Levy entraîne que le critère de Thurston est satisfait dans ce cas. Si on suppose seulement les points critiques de P et Q prépériodiques, il y a des difficultés, peut-être purement formelles.

Notons $C(d)$ l'ensemble des polynômes moniques centrés de degré d dont l'ensemble de Julia rempli est connexe.

QUESTION. *Le polynôme P satisfaisant aux hypothèses ci-dessus, et un point de base fixe par P^k étant choisi sur ∂U_α , existe-t-il une application continue injective de $C(\delta)$ dans $C(d)$ qui à Q associe un modulé de P par Q lorsque $K(Q)$ est localement connexe?*

On conjecture une réponse positive pour $\delta = 2$. Mais on ne peut évidemment pas utiliser la méthode qui réussit lorsque $d = 2$, qui repose sur une connaissance détaillée de la combinatoire de M . Pour $\delta > 2$, il est raisonnable de penser que les phénomènes qui créent des discontinuités pour le redressement des applications à allure polynomiale se produisent aussi, entraînant une réponse négative.

On peut faire des variantes. Si P admet deux cycles disjoints superattractifs de degré 2, i.e., contenant chacun un point critique simple, on peut chercher à les moduler indépendamment. Mais, pour certains P , on n'obtiendra au mieux qu'une application de $M^2 - \{(\frac{1}{4}, \frac{1}{4})\}$ dans $C(d)$. Par exemple, Milnor a regardé la partie réelle $\mathbf{RC}(3)$ de $C(3)$, et on observe le phénomène suivant: Le polynôme

$z \rightarrow z^3 + \frac{3}{2} \cdot z$ a deux points fixes superattractifs imaginaires purs conjugués. Si on les module l'un par c , l'autre par \bar{c} avec $c \in M - \{\frac{1}{4}\}$, on obtient un polynôme réel. Mais l'application de $M - \{\frac{1}{4}\}$ dans $\mathbf{RC}(3)$ ainsi définie étale le point de rebroussement de la cardioïde sur la courbe Γ formée des polynômes ayant un point fixe double. Les images des "éléphants" qui se trouvent près de ce point de rebroussement s'accumulent sur un arc de Γ , leur hauteur tendant vers une limite strictement positive tandis que leur largeur tend vers 0. En particulier, $\mathbf{RC}(3)$ n'est pas localement connexe. Ces faits, observés expérimentalement, peuvent se justifier en utilisant les "cylindres d'Ecalte." Remarquons qu'on sait (Branner-Hubbard) que $C(3)$ et $\mathbf{RC}(3)$ sont des compacts connexes, et même cellulaires (intersections décroissantes de boules topologiques).

Le polynôme $z \rightarrow z^3 - \frac{3}{2} \cdot z$ a un cycle superattractif contenant deux points critiques. Les objets par lesquels on peut espérer le moduler sont les systèmes dynamiques formés de deux copies de \mathbf{C} , chacune étant appliquée dans l'autre par un polynôme de degré 2. Pour un tel système dynamique $Q = (Q_1, Q_2)$, on pose $K_1(Q) = K(Q_2 \circ Q_1)$ et $K_2(Q) = K(Q_1 \circ Q_2)$. Ces deux ensembles sont connexes si aucun des deux points critiques ne s'échappe vers l'infini. Sinon, ils ont chacun une infinité de composantes connexes. On peut donc s'attendre à trouver dans $C(3)$ une copie de l'ensemble $C(2, 2)$ des couples (c_1, c_2) tels que les $K_i(z^2 + c_1, z^2 + c_2)$ soient connexes.

Accouplements. Soient P_1 et P_2 deux polynômes moniques de même degré d , et posons $K_i = K(P_i)$. Supposons K_1 et K_2 connexes et localement connexes, et notons γ_i le lacet de Caratheodory de K_i . On définit un espace topologique Σ en identifiant dans la réunion disjointe de K_1 et K_2 les points $\gamma_1(t)$ et $\gamma_2(-t)$. On a une application $F: \Sigma \rightarrow \Sigma$ induisant P_1 sur K_1 et P_2 sur K_2 . Si Σ est homéomorphe à la sphère S^2 et F topologiquement conjuguée à une fraction rationnelle f , avec une conjuguante holomorphe aux points intérieurs à l'un des K_i , on dit que f réalise l'accouplement de P_1 et P_2 .

On a de nombreux exemples où ce miracle se produit. Soit f une fraction rationnelle de degré d à ensemble postcritique fini. S'il existe une courbe simple γ évitant l'ensemble postcritique de f , telle que $f^{-1}(\gamma)$ soit une courbe simple isotope à γ modulo l'ensemble postcritique, alors f réalise l'accouplement de deux polynômes de degré d .

On peut modifier cette situation de façon à autoriser dans certains cas la courbe γ à passer par certains points post-critiques. Ainsi l'application de Lattès

$$z \rightarrow (i/2) \cdot (z + 1/z),$$

déduite de la multiplication complexe par $1+i$ dans $\mathbf{C}/(\mathbf{Z} + \mathbf{Z}i)$, réalise l'accouplement de deux copies du polynôme $z \rightarrow z^2 + c$ où c est le point de M d'argument externe $\frac{1}{4}$. Remarquons qu'ici les deux K_i sont d'intérieur vide, et leur lacet de Caratheodory commun décrit une courbe de Peano dans la sphère de Riemann. On ne voit pas du tout d'où vient la structure complexe sur Σ . Elle est cependant unique.

Il arrive souvent qu'une même fraction rationnelle réalise plusieurs accouplements.

Partons maintenant de deux polynômes moniques P_1 et P_2 donnés, et cherchons à savoir s'ils sont accouplables. Si tous les points critiques de P_1 et P_2 sont périodiques, la question relève en principe du théorème de Thurston mentionné plus haut. On peut définir l'*accouplement topologique* de la façon suivante: Choisissons un potentiel η et notons $L(P_i)$ l'ensemble des points où le potentiel par rapport à $K(P_i)$ est $< \eta$. L'ouvert $A_i = L(P_i) - K(P_i)$ est alors un anneau de module $\eta/2\pi$. En recollant $L(P_1)$ et $L(P_2)$ suivant l'isomorphisme de A_1 sur A_2 donné par $s \mapsto \eta - s$ pour le potentiel et $t \mapsto -t$ pour l'argument, on obtient une sphère de Riemann X_η . L'application $F_\eta: X_\eta \rightarrow X_\eta$ qui induit P_i sur $K(P_i)$, et qui sur l'anneau $A = A_1 = A_2$ double l'argument sans changer le potentiel, est un revêtement ramifié à points critiques périodiques. Sa classe de Thurston ne dépend pas du choix de η : c'est l'*accouplement topologique*.

THÉOREME. *Supposons que F_η soit équivalente au sens de Thurston à une application rationnelle f . Alors*

- (a) P_1 et P_2 sont accouplables, et f réalise l'accouplement;
- (b) en choisissant pour chaque η un isomorphisme $\varphi_\eta: X_\eta \rightarrow \overline{\mathbb{C}}$ de façon appropriée, $f_\eta = \varphi_\eta \circ F_\eta \circ \varphi_\eta^{-1}$ tend vers f quand η tend vers 0;
- (c) avec ces choix, la restriction de φ_η à $K(P_i)$ a une limite ι_i (non nécessairement injective) quand η tend vers 0.

De fait, la construction de Thurston transforme f_η en $f_{\eta/2}$.

Cependant le critère de Thurston est très difficile à rendre explicite. Pour les polynômes de degré 2, on a une condition nécessaire conjecturalement suffisante. Notons W_0 la grande cardiïde de M . La représentation conforme de W_0 permet de définir pour chaque point de ∂W_0 un *argument interne*. Pour chaque $t \in \mathbb{Q}/\mathbb{Z}$, notons M_t le *membre* de M d'argument interne t , i.e., l'ensemble des points de M qu'on ne peut joindre à 0 par un ensemble connexe qu'en passant par le point de W_0 d'argument interne t . L'ensemble M est réunion de \overline{W} et des M_t . Soient c_1 et c_2 deux points de M tels que 0 soit périodique pour P_{c_1} . Si c_1 et c_2 sont dans des membres de M conjugués (i.e. symétriques par rapport à \mathbb{R}), ils ne sont pas accouplables. On conjecture que dans tous les autres cas ils le sont. Traduisant le critère de Thurston, Silvio Levy a réduit le problème de l'accouplabilité à l'impossibilité de trouver une courbe ayant certaine propriété combinatoire. Tan Lei a alors obtenu le résultat suivant: supposons c_1 et c_2 dans M_{t_1} et M_{t_2} respectivement, avec $t_1 + t_2 \neq 1$. Supposons en outre que c_1 soit le modulé d'un point c'_1 de la nervure principale de M_{t_1} par un point c_1 , et que ni t_1 ni t_2 ne soit de la forme $1/n$ ou $-1/n$. Alors c_1 et c_2 sont accouplables.¹

La traduction du théorème de Thurston pour les polynômes à ensemble postcritique fini présente quelque complication formelle.

Pour l'extension aux polynômes qui ne sont pas à ensemble postcritique fini, ou pour la continuité en fonction des données, il n'y a aucun résultat démontré.

¹ *Added in proof:* Mary Rees a obtenu la conjecture complète.

Mais il y a une très belle étude expérimentale de Ben Wittner sur les fractions rationnelles de degré 2 ayant un cycle superattractif d'ordre 3. Ces fractions rationnelles forment une famille à 1 paramètre, et dans le plan du paramètre on voit l'image de M mutilé d'un membre, correspondant aux accouplements avec un lapin, le symétrique et on voit aussi les points correspondant aux accouplements avec le point réel $-1,754877$ de période 3. Ben Wittner a des algorithmes, basés sur une mise en oeuvre de la méthode de Thurston, qui lui permettent de tracer les nervures de M et leurs images par accouplement avec un polynôme donné (à point critique périodique). Cela suggère très fortement que ces applications sont continues (encore qu'avec un assez mauvais module de continuité). On peut repérer les fractions rationnelles qui peuvent être obtenues par accouplement de plusieurs façons, et celles qui, bien qu'ayant deux cycles superattractifs, ne peuvent pas être obtenues par accouplement.

Les inégalités de Shishikura. Dans un article récent, Mitsuhiro Shishikura utilise une méthode de chirurgie pour démontrer les conjectures qui circulaient sur le nombre maximum de cycles non répulsifs ou de cycles de composantes du complémentaire de l'ensemble de Julia d'une fraction rationnelle de degré donné. Pour ce faire, il a affiné l'outil: il obtient le lemme suivant:

LEMME. *Soit $g: \overline{C} \rightarrow \overline{C}$ une application quasi-régulière. Soient E_i des ouverts disjoints de \overline{C} ($i = 1, \dots, m$), soient $\phi_i: E_i \rightarrow E'_i$ des applications quasi-conformes, où les E'_i sont des ouverts de \overline{C} , et soit N un entier. On suppose satisfaites les conditions suivantes:*

- (i) $g(E)$ est contenu dans E , ou $E = \bigcup E_i$;
- (ii) $\phi \circ g \circ (\phi_i^{-1})$ est analytique sur E'_i , ou $\phi: E \rightarrow \overline{C}$ a pour restriction les E_i ;
- (iii) $\partial g / \partial \bar{z} = 0$ presque partout sur $\overline{C} - g^{-N}(E)$.

Il existe alors une application quasi-conforme $\phi: \overline{C} \rightarrow \overline{C}$ telle que $\phi \circ g \circ \phi^{-1}$ soit une fraction rationnelle. De plus, $\phi \circ \phi_i^{-1}$ est conforme sur E'_i , et $\partial \phi / \partial \bar{z} = 0$ presque partout sur $\overline{C} - \bigcup g^{-N}(E)$.

La démonstration est une adaptation de celle du théorème de redressement des applications à allure polynomiale.

Une application rationnelle f de degré d étant donnée, notons n_a le nombre de cycles attractifs (en comptant les superattractifs), n_p le nombre de cycles paraboliques (i.e., indifférents rationnels), n_P le nombre de cycles de domaines paraboliques (on a $n_P \geq n_p$ car plusieurs cycles de domaines paraboliques peuvent être attachés à un même cycle de points), n_{irr} le nombre de cycles indifférents irrationnels, et n_H le nombre d'anneaux de Herman. Quand il y a lieu de préciser, on écrira $n_a(f)$, etc.

THÉOREME (SHISHIKURA). *Soit f une application rationnelle de degré d . On a $n_a + n_P + n_{irr} + 2 \cdot n_H \leq 2d - 2$.*

Commençons par un résultat intermédiaire. Pour un polynôme, notons n_a^* le nombre de cycles attractifs à distance finie, soit $n_a^* = n_a - 1$.

- PROPOSITION. (a) Pour un polynôme de degré d , on a $n_a^* + n_P + n_{\text{irr}} \leq d - 1$.
 (b) Pour une fraction rationnelle de degré d , on a $n_a + n_P + n_{\text{irr}} \leq 2d - 2$.

INDICATION DE DÉMONSTRATION. (a) Le résultat était connu [7], il découle de la notion même d'application à allure polynomiale. Comme chaque bassin attractif ou parabolique contient au moins un point critique, on a $n_a + n_P \leq d - 1$. Il n'est pas évident que l'on peut perturber un polynôme f , sans changer son degré, de façon à rendre les points indifférents irrationnels attractifs et à conserver le nombre des bassins paraboliques. Mais cela est facile dans le cadre des applications à allure polynomiale: soit $f: U' \rightarrow U$ une application à allure polynomiale de degré d , et choisissons un polynôme h (de degré élevé) tel que:

- (i) h s'annule aux points périodiques non répulsifs de f ;
- (ii) $\operatorname{Re}(h'(x)/f'(x)) < 0$ si x est un point périodique indifférent irrationnel;
- (iii) Si x est un point parabolique, h s'annule en x à un ordre suffisant pour que le nombre de pétales de la fleur de Fatou en x soit le même pour $f + \varepsilon h$ et pour f .

Soit U_1 l'ouvert U légèrement rétréci. Pour ε assez petit, $f + \varepsilon h$ induit une application à allure polynomiale $f_\varepsilon: U'_\varepsilon \rightarrow U_1$, vérifiant: $n_a(f_\varepsilon) \geq n_a(f) + n_{\text{irr}}(f)$, $n_P(f_\varepsilon) \geq n_P(f)$. On a donc $n_a(f) + n_{\text{irr}}(f) + n_P(f) \leq n_a(f_\varepsilon) + n_P(f_\varepsilon) \leq d - 1$.

(b) Soit f une fraction rationnelle telle que $n_{\text{irr}}(f) > 0$. On peut supposer que 0 est un point périodique indifférent irrationnel, que $f(\infty) = 0$, mais que ∞ n'appartient pas au cycle de 0. Choisissons un polynôme h satisfaisant aux conditions (i), (ii), (iii) énoncées plus haut. Pour $\varepsilon > 0$, notons V_ε le bassin immédiat de 0 pour ε , et $\rho(\varepsilon)$ la distance de 0 à ∂V_ε . Si 0 est une point de Siegel, $\rho(\varepsilon)$ ne tend pas vers 0; sinon, il tend vers 0, mais plus lentement que ε^β pour tout β .

D'autre part, on peut choisir un disque D_ε , de rayon R^ε tendant vers l'infini comme $1/\varepsilon^\beta$ pour un certain β , de façon que f_ε ressemble à f sur D_ε . Pour ε assez petit, $f_\varepsilon(\partial D_\varepsilon)$ est donc contenu dans V_ε . On peut alors raccorder f_ε à f sur le complémentaire de D_ε . On obtient une application qui n'est plus holomorphe, mais qui peut le devenir si on change la structure complexe de la sphère de Riemann. Pour cette dernière application g , on a $n_a(g) \geq n_a(f) + n_{\text{irr}}(f)$, et $n_P(g) \geq n_P(f)$. D'où l'inégalité cherchée.

Shishikura et les anneaux de Herman. Pour terminer la démonstration de son théorème, Shishikura doit traiter le cas où il y a des anneaux de Herman. Indiquons quelques constructions pour se faire la main.

Comment obtenir un anneau de Herman en accouplant des disques de Siegel:

Soient f_1 et f_2 des applications rationnelles de degré d_1, d_2 , ayant des disques de Siegel invariants D_1, D_2 , de nombre de rotation t_1 et $t_2 = -t_1$. Nous allons construire à partir de f_1 et f_2 , une application rationnelle f_3 ayant un anneau de Herman de nombre de rotation t_1 ou t_2 suivant l'orientation choisie.

Soient γ_1 et γ_2 des cercles invariants dans D_1 et D_2 , et notons D'_1 et D'_2 les sous-disques de D_1 et D_2 limites par γ_1 et γ_2 . On peut construire un difféomorphisme ψ de \overline{C} ayant les propriétés suivantes:

- (i) $\psi(\gamma_1) = \gamma_2$, $\psi \circ f_1 = f_2 \circ \psi$ sur γ_1 , $\psi(D'_1) = \overline{C} - \overline{D'_2}$;
- (ii) Le support de $\partial\psi$ est contenu dans $A = D_1 \cap \psi^{-1}(D_2)$.

Définissons alors g par $g = f_1$ sur $C - D'_1$ et $g = \psi^{-1} \circ f_2 \circ \psi$ sur $\overline{D_1}$. L'application g a les propriétés suivantes:

- $g(A) = A$;
- g est holomorphe sur un voisinage de $\overline{C} - A$;
- La structure presque complexe σ_A que coïncide avec la structure standard σ_0 sur \overline{C} et avec $\psi^*(\sigma_0)$ sur $A \cap D'_1$ est invariante par g .

On peut alors étendre σ_A en une structure presque complexe σ sur \overline{C} invariante par g . Par Ahlfors-Bers, il existe un homéomorphisme quasi-conforme ϕ de \overline{C} qui est holomorphe de σ vers σ_0 . Alors $f_3 = \phi \circ g \circ \phi^{-1}$ est une application rationnelle, admettant $\phi(A)$ comme anneau de Herman.

REMARQUES. (1) En comptant les points critiques, on trouve $d_3 - 1 = (d_1 - 1) + (d_2 - 1)$, soit $d_3 = d_1 + d_2 - 1$.

(2) L'application f_3 n'est pas déterminée par f_1 et f_2 : un anneau de Herman comporte un module.

Comment découper un anneau de Herman invariant en disque de Siegel.

Soit f une application rationnelle de degré d avec un anneau de Herman A . On suppose que $f(A) = A$, et que f opère sur A avec un angle de rotation t (ou $-t$ suivant l'orientation choisie). Choisissons un cercle invariant γ dans A , qui coupe \overline{C} en deux pièces B_1 et B_2 . On peut construire un difféomorphisme ψ de \overline{C} tel que:

- (i) $\psi(\gamma) = S^1$ (cercle unité), et ψ conjugue f sur γ à la rotation d'angle t ;
- (ii) le support de $\partial\psi$ est contenu dans A .

Pour $i = 1, 2$, soit $g_i: \overline{C} \rightarrow \overline{C}$ l'application qui coïncide avec $\psi \circ f \circ \psi^{-1}$ sur $\psi(B_i)$ et avec la rotation d'angle t sur le complémentaire. Soit σ_i^A la structure complexe qui coïncide avec $\psi_*\sigma_0$ sur $\psi(A \cap B_i)$ et avec σ_0 sur le complémentaire dans A . Alors σ_i^A est une structure presque complexe sur $\psi(A)$ invariante par g . On peut l'étendre en une structure presque complexe σ_i sur \overline{C} invariante par g_i avec rapport de dilatation borné puisque g est holomorphe en dehors de $\psi(A)$. On peut alors trouver pour $i = 1, 2$ une application quasi-conforme ϕ_i holomorphe de σ_i vers σ_0 , et $f_i = \phi_i \circ g_i \circ \phi_i^{-1}$ est une application rationnelle ayant un disque de Siegel de nombre de rotation t ou $-t$. On peut reconstruire f à partir de f_1 et f_2 comme il a été expliqué plus haut.

Et s'il y a plusieurs anneaux de Herman...

Soit f une fraction rationnelle de degré d avec des anneaux de Herman périodiques $(A_i)_{i \in I}$. Choisissons dans chaque A_i une courbe périodique γ_i de façon que $f(\gamma_i) = \gamma_{f_*(i)}$, et notons Γ la réunion des γ_i . Chaque anneau A est coupé par γ_i en deux demi anneaux A'_i et A''_i avec $f(A'_i) = A''_i$. Appelons n -pièces les composantes connexes de $\overline{C} - f^{-n}(\Gamma)$. Appelons objets les demi-anneaux ou les points périodiques non repulsifs. Nous dirons que deux objets sont séparés au

niveau n s'ils sont dans deux n -pièces différentes. On peut trouver un niveau N qui sépare tout ce qui est destiné à être séparé. Soit $(B_j)_{j \in J}$ la famille des N -pièces. On peut écrire $J = J' \cup J''$; les pièces B_j , $j \in J'$, sont celles qui contiennent un objet. Pour $j \in J'$, notons $B_{f_*(j)}$ la N -pièce qui contient les images des objets de B_j . On prend la réunion disjointe des B_j pour $j \in J'$, et on colle à chacun d'eux le nombre de disques nécessaires pour en faire une sphère topologique S_j . On peut définir des applications $g_j: S_j \rightarrow S_{f_*(j)}$ de façon à ce que g_j coïncide avec f sur $B_j \cap f^{-1}(B_{f_*(j)})$. En changeant la structure complexe des S_i comme on l'a fait plus haut, on transforme les g_j en des applications rationnelles f_j .

Les B_j pour $j \in J''$, on les oublie. On se retrouve finalement avec un système dynamique holomorphe où l'espace sous-jacent est une réunion disjointe de sphères de Riemann indexée par J' . À un tel système dynamique F , on peut appliquer les théorèmes classiques de Fatou, Julia, Sullivan, etc. sur les applications rationnelles. Également la méthode de perturbation de Shishikura pour rendre les cycles indifférents irrationnels attractifs. En notant $\chi(F)$ le nombre de points critiques de F , on peut étendre la Proposition.

DEMONSTRATION DU THÉORÈME DE SHISHIKURA. Soit f une application rationnelle de degré d , et soit F le système dynamique construit à partir de f . Chaque anneau de Herman est remplacé par deux disques de Siegel, les points périodiques et les bassins paraboliques sont conservés, de sorte que l'on a:

$$n_a(F) \geq n_a(f), \quad n_P(F) \geq n_P(f), \quad n_{\text{irr}}(F) \geq n_{\text{irr}}(f) + 2 \cdot n_H(F).$$

On peut perturber F en un système dynamique d'applications rationnelles F_ε de façon que les cycles indifférents irrationnels deviennent attractifs, sans perturber les bassins paraboliques. On a alors $n_a(F_\varepsilon) \geq n_a(F) + n_{\text{irr}}(F)$. D'où

$$\begin{aligned} n_a(f) + n_P(f) + n_{\text{irr}}(f) + 2 \cdot n_H(f) &\leq n_a(F) + n_P(F) + n_{\text{irr}}(F) \\ &\leq n_a(F_\varepsilon) + n_P(F_\varepsilon) \leq \chi(F_\varepsilon) \\ &= \chi(F) \leq \chi(f) = 2d - 2. \end{aligned}$$

Ceci achève la démonstration du Théorème.

Copies d'un membre de M dans des espaces de paramètres. Notons $M_{1/2}$ le membre de l'ensemble de Mandelbrot M d'argument interne $\frac{1}{2}$, i.e., la partie de M située à gauche du point $-\frac{3}{4}$. Partant d'une idée de J. C. Yoccoz, Bodil Branner obtient un homéomorphisme de $M_{1/2}$ sur l'ensemble des polynômes moniques centrés de degré 3 tels que l'image d'un des points critiques soit le point fixe d'argument externe 0. Nous esquissons la construction:

Partons d'un polynôme quadratique P avec $c \in M_{1/2}$. Il y a deux points fixes: α d'arguments externes $\frac{1}{3}$ et $\frac{2}{3}$, et β d'argument externe 0. Le point $\alpha' = -\alpha$ a pour arguments externes $\frac{1}{6}$ et $\frac{5}{6}$. Fixons un potentiel s et notons V , V' et V'' les compacts limités par les équipotentiels de niveau s , $\frac{s}{2}$ et $\frac{s}{4}$. Les lignes

$$L = R_{1/3} \cup R_{2/3} \cup \{\alpha\} \quad \text{et} \quad L' = R_{1/6} \cup R_{5/6} \cup \{\alpha'\}$$

coupent V en trois pièces V_0 , V_1 et V_{-1} , avec $0 \in V_0$, $-\beta \in V_1$ et $\beta \in V_{-1}$. Posons $V'_0 = V' \cap V_0$, etc.

On définit un espace X en identifiant dans $V_0 \cup V_{-1}$ les rayons $R_{1/3}$ et $R_{2/3}$, potentiel par potentiel. Appliquons V'_{-1} dans X par P_c et V'_0 dans X par P_c^2 . On a une discontinuité le long de L' . En la lissant, on obtient une application $g: X' \rightarrow X$ qui est un revêtement ramifié de degré 3, avec X' contenu dans l'intérieur de X . Si on s'y est bien pris, on peut trouver sur X une structure presque complexe invariante par g , à rapport de dilatation borné. On l'intègre et on obtient une application à allure polynomiale de degré 3. En la redressant, on obtient un polynôme cubique Q_c .

La continuité de $c \mapsto Q_c$ n'est pas évidente, mais on la démontre en utilisant de façon répétée la méthode qui réussit pour le redressement en degré 2. Pour montrer l'injectivité, on fait la construction dans l'autre sens.

Dans le même ordre d'idées, et en m'inspirant de la technique de Shishikura, je construis une application continue de $M_{1/2}$ dans le membre $M_{1/3}$, appliquant la racine sur la racine, et $-2 = \gamma_M(\frac{1}{2})$ sur $\gamma_M(\frac{1}{4})$. Cela montre que la *nervure* de M qui mène à $\gamma_M(\frac{1}{4})$ est effectivement un arc (un pas en direction de la connexité locale).

Partant d'un polynome P_c comme ci-dessus, on définit un espace Y en prenant $V_0 \cup V_{-1}$ et deux copies $V_1^{(1)}$ et $V_1^{(2)}$ de V_1 , et en cousant $V_1^{(1)}$ à V_0 suivant $R_{1/3}$, $V_1^{(2)}$ à V_0 suivant $R_{2/3}$, et $V_1^{(1)}$ à $V_1^{(2)}$ par P_c sur $R_{1/3} \cap V_1'$.

On applique V'_0 dans $V_1^{(1)}$ par P_c , $V_1^{(1)}$ dans $V_1^{(2)}$ par l'identité, $V_1'^{(2)}$ sur $V_0 \cup V_{-1}$ par P_c , et aussi V'_{-1} sur $V_0 \cup V_{-1}$ par P_c . On lisse les discontinuités et on obtient une application $g: Y' \rightarrow Y$ qui est un revêtement ramifié. On construit une structure complexe invariante par g avec rapport de dilatation bornée, ce qui est un peu plus délicat que dans le cas précédent, et on termine de même.

REFERENCES

1. P. Blanchard, *Complex analytic dynamics on the Riemann sphere*, Bull. Amer. Math. Soc. (N.S.) 11 (1984), 85-141.
2. B. Branner et A. Douady, *Surgery on complex polynomials*, Proceedings of the Symposium on Dynamical Systems (Mexico, 1986), Preprint D. T. H., Lyngby, Danemark (to appear).
3. A. Douady, *Systèmes dynamiques holomorphes*, Bourbaki seminar, Vol. 1982/83, Astérisque, 105-106, Soc. Math. France, Paris, 1983, pp. 39-63.
4. —, *Veins of the Mandelbrot set*, MSRI, 1985.
5. —, *Algorithms for computing angles in the Mandelbrot set*, Chaotic Dynamics and Fractals (Barnsley-Demko), Proceedings of the Conference on Chaotic Dynamics (Georgia Tech., April 1985), Academic Press, New York.
6. A. Douady et J. H. Hubbard, *Itération des polynômes complexes*, Publ. Math. Orsay, 1984-02, 1985-04.
7. —, *On the dynamics of polynomial-like mappings*, Ann. Sci. École Norm. Sup. (4) 18 (1985).
8. —, *A proof of Thurston topological characterization of rational functions*, Preprint, Inst. Mittag-Leffler, Stockholm, 1985.
9. Tan Lei, *Accouplement des polynômes quadratiques complexes*, Note C.R. Acad. Sci. Paris, February 1986.

10. S. V. F. Levy, *Critically finite rational maps*, Thèse Ph.D., Princeton Univ., 1985.
11. M. Shishikura, *On the quasi-conformal surgery of rational functions*, Preprint, Kyoto, 1986 (to appear in Ann. Sci. École Norm. Sup.).
12. D. Sullivan, *Quasi-conformal homeomorphisms and dynamics*, Preprint, I.H.E.S., Bures-sur-Yvette, 1982.
13. W. P. Thurston, notes intimes.

UNIVERSITÉ DE PARIS-SUD, 91405 ORSAY, FRANCE

ECOLE NORMALE SUPÉRIEURE, 75005 PARIS, FRANCE

Рациональные аппроксимации аналитических функций

А. А. ГОНЧАР

В статье излагаются результаты последних лет, относящиеся к задаче о скорости рациональной аппроксимации аналитических функций.

1. Пусть E — компакт в расширенной комплексной плоскости $\hat{\mathbb{C}}$, f — функция, непрерывная на E , \mathcal{R}_n класс всех рациональных функций от z порядка не выше n . Уклонение f от \mathcal{R}_n (в равномерной метрике на E) обозначим через $r_n = r_n(f, E)$:

$$r_n = \inf\{\|f - r\|_E : r \in \mathcal{R}_n\},$$

где $\|\cdot\|_E$ — \sup — норма на E . Если функция f голоморфна на компакте E ($f \in H(E)$), то последовательность r_n стремится к нулю геометрически, точнее,

$$\overline{\lim}_n r_n^{1/n} = q < 1.$$

Величина $q = q(f, E)$ является основной характеристикой скорости рациональной аппроксимации f на E (во всяком случае, при $q \in (0, 1)$). Вопрос о вычислении (характеризации) этой величины в терминах, связанных с аналитическим продолжением функции f , играет фундаментальную роль в теории рациональных аппроксимаций аналитических функций.

Всюду в дальнейшем будем предполагать, что компакт E состоит из конечного числа нетривиальных связных компонент (континуумов). Пусть F — компакт, принадлежащий $\hat{\mathbb{C}} \setminus E$, $h(E, F)$ — модуль конденсатора (E, F) ($h = 1/c$, где $c = c(E, F)$ — емкость этого конденсатора). Из результатов Дж. Уолша [15] и Т. Бегби [1], относящихся к интерполяции рациональными функциями с фиксированными полюсами, вытекает следующая теорема: если функция f голоморфна в открытом множестве D , $E \subset D = \hat{\mathbb{C}} \setminus F$, то для $f(z)$, $z \in E$, справедлива оценка

$$q \leq \exp(-h(E, F)) \quad (1)$$

Обозначим через \mathcal{F} класс всех компактов F , таких, что функция $f \in H(E)$ допускает голоморфное (однозначное аналитическое) продолжение в $\hat{C} \setminus F$. Модулем голоморфности функции $f \in H(E)$ назовем величину

$$h = \sup\{h(E, F) : F \in \mathcal{F}\} \quad (0 < h \leq +\infty).$$

В наиболее интересных случаях существует единственный регулярный компакт $F_f \in \mathcal{F}$, для которого $h(E, F_f) = h$; открытое множество $D_f = \hat{C} \setminus F_f$ — максимальная область голоморфности функции $f \in H(E)$ (в плоскости \hat{C}). Из (1) вытекает, что для любой функции $f \in H(E)$ величины q и h связаны неравенством

$$q \leq e^{-h}. \quad (2)$$

Отметим, что обе введенные характеристики (q и h) инвариантны относительно дробно-линейных преобразований плоскости \hat{C} .

Оценка (2) позволяет вычислить q только при $h = +\infty$. Последнее равенство означает, что функция $f \in H(E)$ допускает голоморфное продолжение в область вида $\hat{C} \setminus F$, $\text{cap } F = 0$ (cap — логарифмическая емкость); для таких функций

$$q = \lim_n r_n^{1/n} = 0.$$

В общем случае величину q нельзя охарактеризовать в терминах, связанных только с аналитическим продолжением функции f ; в частности, q нельзя выразить через h . Хорошо известны также примеры функций, для которых $0 < q = e^{-h}$. Однако, соответствующие функции имеют довольно экзотический характер. В последние годы получен ряд результатов, показывающих, что для широких классов аналитических функций, включающих важнейшие функции анализа, q можно выразить через h ; при этом, имеет место равенство

$$q = \lim_n r_n^{1/n} = e^{-2h}. \quad (3)$$

По-видимому, первые результаты общего характера, приводящие к формуле (3), были получены в работах автора [2, 3]. Для доказательства соответствующих результатов в этих работах был применен метод, основанный на интерполяции функции $f \in H(E)$ рациональными функциями со свободными полюсами (метод многоточечных аппроксимаций Паде). В [2] рассмотрен случай, когда E и F — отрезки вещественной прямой \mathbb{R} и

$$f(z) = \int_F \frac{d\sigma(t)}{z - t}, \quad z \in E,$$

где σ — положительная мера, носитель которой принадлежит отрезку F . Основной результат этой работы заключается в том, что при условии: $\sigma' = d\sigma/dt > 0$ почти всюду на F имеет место формула

$$q = \exp(-2h(E, F)).$$

Модуль $h(E, F)$ выражается через полные эллиптические интегралы первого рода; если $E = [a, b]$, $F = [c, d]$, то

$$h(E, F) = \pi K' / K, \quad k^2 = \frac{(a-b)(c-d)}{(d-b)(c-a)}.$$

В работе [3] формула (4) доказана для случая, когда E — компакт, симметричный относительно \mathbf{R} , F — объединение конечного числа отрезков этой прямой и f — функция, голоморфная в области $D = \hat{\mathbf{C}} \setminus F$ и имеющая достаточно правильный скачок на F . Из основной теоремы этой работы следует, что равенство (3) справедливо для любой функции $f \in H(E)$, допускающей аналитическое продолжение вдоль путей, принадлежащих области вида $\hat{\mathbf{C}} \setminus e$, где e — компакт нулевой емкости, $e \subset \mathbf{R}$.

Остановимся на схеме применения многоточечных аппроксимаций Паде к задаче о скорости равномерной аппроксимации аналитических функций; подробнее см. [3]. Пусть $f \in H(E)$, $\alpha = \{\alpha_{n,k}\}$, $k = 1, 2, \dots, n$, $n = 1, 2, \dots$ — треугольная таблица точек (узлов интерполяции), принадлежащих компакт E , и $A_n(z) = (z - \alpha_{n,1}) \cdots (z - \alpha_{n,n})$. Фиксируем натуральное n и рассмотрим рациональную функцию $R_n = P_n/Q_n$, где P_n, Q_n — произвольные полиномы от z , удовлетворяющие условиям

$$\deg P_n \leq n-1, \quad \deg Q_n \leq n \quad (Q_n \neq 0); \quad \frac{Q_n f - P_n}{A_{2n}} \in H(E). \quad (5)$$

Последнее требование означает, что $Q_n f - P_n = 0$ в точках $2n$ -ой строки таблицы α (с учетом их кратностей). Полиномы P_n, Q_n , удовлетворяющие условиям (5), существуют для любой $f \in H(E)$; их отношение P_n/Q_n определяет единственную рациональную функцию R_n (с точностью до обычного отождествления). Эта рациональная функция называется многоточечной аппроксимацией Паде (типа $[n-1/n]$) функции f , соответствующей таблице узлов интерполяции α . Если $Q_n \neq 0$ в узлах $\alpha_{2n,1}, \dots, \alpha_{2n,2n}$, то R_n интерполирует f в этих точках. В противном случае при переходе от $Q_n f - P_n$ к $f - R_n$ часть условий интерполяции теряется; однако, потеря d_n условий интерполяции сопровождается понижением степеней числителя и знаменателя функции R_n на величину d_n . Основные трудности, связанные с применением многоточечных аппроксимаций Паде (соответствующих выбранной таблице узлов интерполяции), заключаются в анализе асимптотического поведения их полюсов и проблемы сходимости самих аппроксимаций. Если функция $f \in H(E)$ допускает голоморфное продолжение в открытое множество $D = \hat{\mathbf{C}} \setminus F$ (без ограничения общности можно считать, что E, F — компакты в \mathbf{C} и $f(\infty) = 0$), то полиномы Q_n удовлетворяют комплексным соотношениям ортогональности:

$$\int_{\gamma} Q_n(t) t^j \frac{f(t) dt}{A_{2n}(t)} = 0, \quad j = 0, 1, \dots, n-1, \quad (6)$$

где γ — произвольный контур, охватывающий F (γ принадлежит дополнению к $E \cup F$); для разности $f - R_n$ имеет место формула

$$(f - R_n)(z) = \frac{1}{2\pi i} \cdot \frac{A_{2n}(z)}{(Q_n Q)(z)} \int_{\gamma} \frac{(Q_n Q f)(t) dt}{A_{2n}(t)(z - t)}, \quad (7)$$

в которой Q — произвольный полином степени не выше n . В том случае, когда F — система гладких контуров (разрезов) в \mathbb{C} , и функция f имеет достаточно правильный скачок на F , соотношения ортогональности (6) можно записать в виде

$$\int_F Q_n(t) t^j \cdot \frac{\chi(t) dt}{A_{2n}(t)} = 0, \quad j = 0, 1, \dots, n-1 \quad (8)$$

(χ — скачок f на F); аналогично можно переписать формулу (7).

Применение соотношений (6) или (8) для анализа предельного распределения нулей полиномов Q_n осложняется двумя обстоятельствами. Во-первых, в наиболее интересных случаях (например, когда элемент $f \in H(E)$ определяет многозначную аналитическую функцию с конечным числом точек ветвления) область голоморфности $D = \mathbb{C} \setminus F$ функции f и, тем самым, система разрезов F , определяются по f неоднозначно; во-вторых, мера $f(t) dt / A_{2n}(t)$, фигурирующая в соотношениях ортогональности для Q_n , зависит от номера этого полинома (эта зависимость связана с выбором таблицы узлов интерполяции). Имеющиеся результаты (частично о них будет говориться ниже) показывают, что нули Q_n (полюсы R_n) «выбирают» систему разрезов, обладающую определенными свойствами «симметрии» относительно E и α — если свойства функции f допускают такой выбор; предельное распределение нулей Q_n на этой системе разрезов F также зависит от свойств таблицы α . Отметим, что многие конкретные проблемы в этом направлении пока остаются открытыми.

Вернемся к рассматриваемой задаче. Пусть $\lambda = \lambda_F - \lambda_E$ — равновесный заряд конденсатора (E, F) (λ_E, λ_F — единичные меры, сосредоточенные на E и F , соответственно), $V = V^\lambda$ — логарифмический потенциал заряда λ :

$$V^\lambda(z) = \int \log \frac{1}{|t - z|} d\lambda(t), \quad z \in \mathbb{C}.$$

В случае, когда E — компакт, симметричный относительно \mathbb{R} , F — система отрезков прямой \mathbb{R} и функция $f \in H(D)$, $D = \mathbb{C} \setminus F$, имеет правильный (например, аналитический) скачок на F , справедливо следующее утверждение:

(*) если точки таблицы α имеют предельное распределение, характеризуемое мерой λ_E , то нули Q_n (полюсы R_n) также имеют предельное распределение и это распределение характеризуется мерой λ_F ; последовательность R_n сходится по емкости к f внутри (на компактных подмножествах) области D , причем

$$|f - R_n|^{1/n} \rightarrow \exp 2(V - h_F) \quad (9)$$

по емкости внутри D ; здесь h_F — значение V на F (так что правая часть (9) равна e^{-2h} , $h = h_F - h_E = h(E, F)$, на E).

Это утверждение играет ключевую роль при применении метода многоточечных аппроксимаций Паде к задаче о скорости равномерной аппроксимации рациональными функциями. В работе [3] показано, что отсюда уже следует формула (4) (доказательство неравенства $q = \overline{\lim} r_n^{1/n} \leq e^{-2h}$ не составляет труда; в [3] доказано, что неравенство $\underline{\lim} r_n^{1/n} \geq e^{-2h}$ также вытекает из (9)). Для случая, когда скачок f на F характеризуется положительной мерой σ , анализ значительно упрощается; в этом случае многочлены Q_n удовлетворяют соотношениям ортогональности вида

$$\int_F Q_n(t) t^j \frac{d\sigma(t)}{A_{2n}(t)} = 0, \quad j = 0, 1, \dots, n-1$$

(с весами, зависящими от номера многочлена). Общая теорема о предельном распределении нулей таких многочленов получена в [6]; для рассмотренного выше случая соответствующий результат содержится в [5].

Проблематика, связанная с предельным распределением полюсов и сходимостью аппроксимаций Паде, получила принципиальное развитие в работах Г. Шталя [10–12]. Основной результат Шталя относится к случаю, когда E — континуум со связным дополнением и f допускает продолжение вдоль любого пути, принадлежащего области вида $\hat{C} \setminus e$, где e — компакт нулевой емкости (интерес представляет случай, когда элемент $f \in H(E)$ определяет многозначную аналитическую функцию в этой области). В этом случае существует единственная область D , $E \subset D = \hat{C} \setminus F$, такая, что $h(E, F) = h$ и D содержит любую другую область голоморфности функции f , обладающую тем же свойством максимальности модуля (h — модуль голоморфности f , $F = F_f$, $D = D_f$). Шталь доказал, что для многоточечных аппроксимаций Паде рассматриваемых функций $f \in H(E)$ справедливо утверждение (*) (с $F = F_f$ и $D = D_f$). Как было отмечено выше, отсюда уже следует формула (4) с $F = F_f$ и, тем самым, формула (3) — для произвольной $f \in H(E)$, определяющей многозначную аналитическую функцию, множество особых точек которой имеет нулевую емкость. Метод Шталя основан на том, что компакт (система разрезов) $F = F_f$ и, тем самым, максимальная область голоморфности $D = D_f$, обладают определенным свойством «симметрии» относительно E . При этом, свойство неограниченной продолжаемости f в $\hat{C} \setminus e$ не существенно; важен лишь тот факт, что f имеет достаточно правильный скачок на F_f . Соответствующую общую теорему удобно формулировать именно как теорему о скорости приближения голоморфных функций в «симметричных» (относительно E) открытых множествах $D = \hat{C} \setminus F$ (см. следующий пункт; в п.2, в отличие от п.1, мы приводим точные определения и формулировки теорем).

2. Точку ζ компакта F будем называть правильной, если существует окрестность $U(\zeta)$ этой точки, пересечение которой с F — простая аналитическая дуга. Множество всех правильных точек компакта F обозначается через F_0 . Через \mathcal{A} будем обозначать множество всех компактов F , таких, что $\text{cap}(F \setminus F_0) = 0$ и $F_0 \neq \emptyset$.

Рассмотрим конденсатор (E, F) ; будем писать $F \in S(E)$ (и говорить, что компакт F и открытое множество $D = \hat{C} \setminus F$ обладают S -свойством относительно E), если $F \in \mathcal{A}$ и

$$\frac{\partial V}{\partial n}(\zeta^1) = \frac{\partial V}{\partial n}(\zeta^2), \quad \forall \zeta \in F_0; \quad (10)$$

здесь V — равновесный потенциал конденсатора (E, F) , ζ^1, ζ^2 — достижимые граничные точки множества D , соответствующие точке $\zeta \in F_0$, $\partial/\partial n$ — производная по нормали к F , внутренней для D .

Общая конструкция компактов $F \in S(E)$ может быть описана следующим образом. Фиксируем E (напомним, что рассматриваются компакты E , являющиеся объединением конечного числа континуумов E_1, \dots, E_N) и компакт $e \subset \hat{C}$ нулевой емкости, не пересекающийся с E . Рассмотрим произвольную риманову поверхность \mathcal{R} , являющуюся двулистным (неразветвленным) накрытием области $\hat{C} \setminus e$; p — соответствующая проекция. Пусть E^1, E^2 — непересекающиеся компакты на \mathcal{R} , такие, что $p(E^j) = E$, $j = 1, 2$; $\omega(\tilde{z})$, $\tilde{z} \in \Omega = \mathcal{R} \setminus (E^1 \cup E^2)$ — гармоническая мера ∂E^1 относительно Ω . Тогда замыкание F множества $p\{\tilde{z} \in \mathcal{R} : \omega(\tilde{z}) = 1/2\}$ принадлежит $S(E)$.

Хорошо известно, что многие экстремальные задачи геометрической теории функций приводят к компактам F и множествам D , обладающим S -свойством.

Пусть $F \in \mathcal{A}$ и f — функция, голоморфная в открытом множестве $D = \hat{C} \setminus F$; будем писать $f \in H_0(D)$, если для любой точки $\zeta \in F_0 \setminus e_0$, где e_0 — компакт нулевой емкости, существуют пределы

$$\lim_{z \rightarrow \zeta^j, z \in D} f(z) = f(\zeta^j), \quad j = 1, 2,$$

причем $f(\zeta^1) \neq f(\zeta^2)$ (как и выше, ζ^1, ζ^2 — достижимые граничные точки D , соответствующие точке ζ ; тем самым, разность $f(\zeta^1) - f(\zeta^2)$ определяет скачок функции f в точке ζ).

ТЕОРЕМА 1. Пусть (E, F) — конденсатор, такой, что $F \in S(E)$, f — функция, принадлежащая $H_0(D)$, $D = \hat{C} \setminus F$. Тогда

$$q = \lim_n r_n^{1/n} = e^{-2h(E, F)}.$$

Выделим два следствия этой теоремы (о первом из них уже говорилось в п.1 в связи с результатами Шталя).

В каждом из следующих случаев справедливо равенство (3):

(i) E — континуум и элемент $f \in H(E)$ определяет многозначную аналитическую функцию с конечным числом особых точек;

(ii) E — объединение попарно непересекающихся континуумов E_1, \dots, E_N и $f(z) = c_j$, $z \in E_j$, $j = 1, \dots, N$ (случай $N = 2$ соответствует классической задаче Золотарева; см [4]).

Теорема 1 сводит решение многих других задач о скорости рациональной аппроксимации аналитических функций к задачам геометрической теории функций, связанным с характеристикой максимальных областей голоморфности (различные варианты проблемы модуля).

Прежде чем формулировать следующую теорему, введем необходимые понятия и обозначения. Пусть (E, F) — конденсатор, $F \in \mathcal{A}$; для простоты будем предполагать, что E, F — компакты в \mathbb{C} и F является объединением конечного числа аналитических жордановых дуг (так, что $F \setminus F_0$ — конечное множество). Фиксируем функцию φ , гармоническую в некоторой окрестности компакта F . Через $M(E, F)$ обозначим множество всех зарядов вида $\mu = \mu_F - \mu_E$, где μ_E, μ_F — единичные меры, сосредоточенные на E и F , соответственно. Можно показать, что существует единственный заряд $\lambda \in M(E, F)$ такой, что

$$\begin{aligned} V^\lambda(z) &\equiv w_1, & z \in E, \\ (V^\lambda + \varphi)(z) &\equiv \min_F (V^\lambda + \varphi) = w_2, & z \in L = \text{supp } \lambda_F; \end{aligned} \quad (11)$$

тем самым, константы w_1, w_2 также однозначно определяются соотношениями равновесия (11). Положим $w = w(E, F, \varphi) = w_2 - w_1$.

Заряд $\lambda = \lambda_F - \lambda_E$ решает проблему равновесия на (E, F) при условии, что пластина F находится во внешнем поле φ . Отметим, что равновесный заряд λ является (единственным) решением следующей задаче о минимуме энергии:

$$I_\varphi(\lambda) = \inf\{I_\varphi(\mu) : \mu \in M(E, F)\},$$

где

$$I_\varphi(\mu) = \int \log \frac{1}{|t-z|} d\mu(t) d\mu(z) + 2 \int \varphi(t) d\mu_F(t).$$

О более общих задачах равновесия при наличии внешних полей см. [7]. В этих задачах задаются компакты (проводники) F_1, \dots, F_m , величины зарядов на каждом из проводников, $m \times m$ -матрица $A = \|a_{jk}\|$, характеризующая закон взаимодействия зарядов, принадлежащих различным проводникам, и внешние поля $\varphi_1, \dots, \varphi_m$, действующие в пределах этих проводников. Решение ряда задач, связанных с рациональными аппроксимациями аналитических функций, формулируется в терминах соответствующих равновесных распределений.

Вернемся к задаче (11). Будем писать $F \in S(E, \varphi)$, если $L = \text{supp } \lambda_F \in \mathcal{A}$ и в каждой точке $\zeta \in L_0$ имеет место равенство

$$\frac{\partial(V^\lambda + \varphi)}{\partial n}(s^1) = \frac{\partial(V^\lambda + \varphi)}{\partial n}(s^2)$$

(смысл обозначений — тот же, что и в (10)).

ТЕОРЕМА 2. Пусть $F \in S(E, \varphi)$, g_n — последовательность функций, голоморфных в окрестности U компакта F , такая, что $(1/2n)\operatorname{Re} g_n \rightrightarrows \varphi$ внутри U . Тогда для последовательности функций

$$f_n(z) = \int_F \frac{e^{-g_n(t)} dt}{z-t}, \quad z \in E,$$

справедливо соотношение

$$\lim_n r_n(f_n, E)^{1/n} = e^{-2w}, \quad (12)$$

где $w = w(E, F, \varphi)$.

Соотношение (12) сохраняется и в несколько более общей ситуации, когда

$$f_n(z) = \int_\gamma \frac{e^{-g_n(t)} dt}{z-t}, \quad z \in E,$$

где $f \in H_0(D)$ и γ — произвольный контур, лежащий в U и охватывающий компакт F (функции $f_n(z)$, $z \in E$, не зависят от γ).

Теорема 2 имеет непосредственное применение к известной задаче о скорости рациональной аппроксимации экспоненты на полуоси. Положим $\rho_n = r_n(e^{-x}, E^+)$, $E^+ = [0, +\infty]$. Подробный обзор результатов типа

$$0 < c_1 \leq \liminf_n \rho_n^{1/n} \leq \overline{\lim}_n \rho_n^{1/n} \leq c_2 < 1$$

и гипотез, связанных с (существованием и) величиной предела $v = \lim \rho_n^{1/n}$ имеется в книге П. Варги [14]. Недавно А. Магнус [8] определил значение константы v с помощью метода Каратеодори-Фейера, приспособленного для рациональных аппроксимаций вещественных функций на отрезке $[-1, 1]$ М. Гуткнехтом и Л. Трефетеном [13]; а именно, $v = \exp(-\pi K'/K)$, где K — полный эллиптический интеграл первого рода, модуль которого определяется из $K(k) = 2E(k)$. Описанная в [8] техника определения величины v в принципиальных пунктах имеет эвристический характер; пока неясно, в частности, как строго обосновать переход от n -кратных интегралов к уравнениям, описывающим предельные распределения (при $n \rightarrow \infty$). Дж. Наттолл использовал такого рода переход для анализа асимптотических свойств полиномов Эрмита-Паде (метод седловой точки Наттолла; см. [9]); обоснование соответствующей техники привело бы к решению ряда задач теории рациональных аппроксимаций, пока остающихся открытыми.

С помощью теоремы 2 можно доказать, что предел $v = \lim \rho_n^{1/n}$ существует, и описать величину v в терминах, связанных с задачей равновесия для конденсатора вида (E^+, F) при условии, что пластина F (аналитическая дуга, симметричная относительно \mathbf{R}) находится во внешнем поле $\varphi(z) = \frac{1}{2} \operatorname{Re} z$, причем $F \in S(E^+, \frac{1}{2} \operatorname{Re} z)$. Эта задача равновесия допускает решение в терминах эллиптических функций и интегралов. На этом пути можно получить следующую — очень простую

— характеризацию величины v : v есть (единственный) положительный корень уравнения

$$\sum_{n=1}^{\infty} a_n x^n = \frac{1}{8}, \quad a_n = \left| \sum_{d|n} (-1)^d d \right|. \quad (13)$$

Если $n = 2^m p_1^{m_1} \dots p_s^{m_s}$ — каноническое представление числа n (p_j — нечетные простые, $m \geq 0$, $m_j \geq 1$), то

$$a_n = |2^{m+1} - 3| \frac{p_1^{m_1+1} - 1}{p_1 - 1} \dots \frac{p_s^{m_s+1} - 1}{p_s - 1}.$$

Вычисление константы v на основе (13) не составляет труда; имеем

$$1/v = 9.2890254919208189187554494359517450610317 \dots$$

(это число с тридцатью знаками приведено в [8]).

На указанном пути можно изучить и более общие задачи, связанные с экспонентой, в частности, охарактеризовать значение

$$v_\theta = \lim_n r_n(e^{-z}, E_\theta), \quad E_\theta = \{z = re^{it} : r \geq 0, |t| \leq \theta < \pi/2\}.$$

Теоремы 1, 2 и результаты, относящиеся к экспоненте, получены автором совместно с Е. А. Рахмановым. Теоремы доказываются методом работы [3], основанным на применении многоточечных аппроксимаций Паде; таблицы узлов интерполяции строятся по мерам λ_E , фигурирующим в соответствующих задачах равновесия. Необходимые асимптотические свойства многоточечных аппроксимаций Паде устанавливаются с помощью надлежащей модификации метода Штала. Различные варианты теоремы 1 получены также Шталем.

ЛИТЕРАТУРА

1. T. Bagby, *On interpolating by rational functions*, Duke Math. J. **36** (1969), 95–104.
2. А. А. Гончар, *О скорости рациональной аппроксимации некоторых аналитических функций*, Мат. сб. **105** (1978), № 1, 147–163.
3. —, *О скорости рациональной аппроксимации аналитических функций*, Труды МИАН **166** (1984), 52–60.
4. —, *О задачах Е. И. Золотарева, связанных с рациональными функциями*, Мат. сб. **78** (1969), № 4, 640–654.
5. А. А. Гончар и Гиермо Л. Лопес, *О теореме Маркова для многоточечных аппроксимаций Паде*, Мат. сб. **105** (1978), № 4, 512–524.
6. А. А. Гончар и Е. А. Рахманов, *Равновесная мера и распределение нулей экстремальных многочленов*, Мат. сб. **125** (1984), № 1, 117–127.
7. —, *О задаче равновесия для векторных потенциалов*, Успехи мат. наук **40** (1985), в. 4, 155–156.
8. A. P. Magnus, *CFGT determination of Varga's constant "1/9"*, Institut Mathématique U.C.L., B-1348, Belgium, 1986, Preprint.
9. J. Nuttall, *Asymptotics of diagonal Hermite-Padé polynomials*, J. Approx. Theory **42** (1984), 299–386.
10. H. Stahl, *The convergence of Padé approximants to functions with branch points*, Tech. Univ. Berlin, 1984, Preprint.
11. —, *On the convergence of generalized Padé approximants*, Tech. Univ. Berlin, 1984, Preprint.

12. —, *A note on three conjectures by Gonchar about rational approximation*, Tech. Univ. Berlin, 1985, Preprint.
13. L. N. Trefethen and M. H. Gutknecht, *The Carathéodory-Fejer method for real rational approximation*, SIAM J. Num. Anal. **20** (1983), 420–436.
14. R. S. Varga, *Topics in polynomial and rational interpolation and approximation*, Univ. Montreal, 1982.
15. ДЖ. Л. Уолш, *Интерполяция и аппроксимация рациональными функциями в комплексной области*, Изд-во иностр. лит., Москва, 1961.

Математический институт им. В. А. Стеклова, Москва 117966, СССР

Description of the Local Automorphism Groups of Real Hypersurfaces

N. G. KRUZHILIN

The study of real hypersurfaces in the space of several complex variables originated with the works of H. Poincaré and E. Levi. Since then, the theory of Levi nondegenerate hypersurfaces has made substantial progress, first of all, due to E. Cartan, H. Tanaka, S. S. Chern, and J. K. Moser. We shall discuss one of the parts of this theory related to the properties of locally defined biholomorphisms and CR-diffeomorphisms of these hypersurfaces and, in particular, to the groups of their local automorphisms. (We use the term local automorphism to designate a holomorphic or CR-transformation defined in some neighborhood of a distinguished point on the hypersurface, leaving this point fixed and mapping the germ of the hypersurface into itself). We shall give a survey of results in this direction obtained in the seminar of A. G. Vitushkin and M. S. Mel'nikov at the Moscow University.

The simplest of all the Levi nondegenerate hypersurfaces are the hyperquadrics in $\mathbf{C}^{n+1}(z^1, \dots, z^n, w = u + iv)$ given by equations

$$v = \sum_{\alpha, \beta} g_{\alpha\beta} z^\alpha \bar{z}^\beta \quad (1)$$

where $(g_{\alpha\beta})$ is a nonsingular Hermitian matrix of signature $(p, n - p)$. As was discovered by Poincaré [1] and later by other authors, locally biholomorphic mappings of one hyperquadric into another are linear fractional. From the viewpoint of the properties of local biholomorphisms, the quadrics are completely investigated.

The general theory developed by E. Cartan [2] for the surfaces in \mathbf{C}^2 and by Chern and Moser [3] for the other dimensions uses the quadrics as a model case. For instance, Moser proved that for every real analytic hypersurface there are holomorphic charts where the equation of the hypersurface has so-called normal form resembling (1). Locally biholomorphic mappings of surfaces given by equations in normal form are close to linear fractional transformations of the corresponding quadrics. The set of these coordinate systems defined in the vicinity of the distinguished point on the surface depends on $(n^2 + 2n + 2)$

real parameters, and there is a correspondence between this set and the local automorphism group of the hyperquadric G . Therefore, the local automorphism group of the hypersurface \mathcal{G} is isomorphic to a subgroup of G .

The further progress in the problems that we consider was connected with introducing the class of nonspherical hypersurfaces, i.e., of hypersurfaces not CR-equivalent to a quadric in any neighborhood of the distinguished point. It turns out that the properties of biholomorphic mappings of nonspherical hypersurfaces differ in many respects from those of the quadrics. On the whole, one may say that nonspherical hypersurfaces are more "inflexible" than quadrics. This "inflexibility" reveals itself in various ways. As examples, we mention results on linearization of local automorphisms and the theorem of A. G. Vitushkin on the germ of a map of pseudoconvex surfaces. The proofs use the normal forms and the chains—a family of curves on Levi nondegenerate hypersurfaces also introduced by Cartan, Chern, and Moser. A special form of the equation of a hypersurface different from that of Moser proved to be useful in the investigations as well.

The first general result on the local properties of maps of nonspherical hypersurfaces is that obtained by V. K. Beloshapka and A. V. Loboda. It relates to parameters of local automorphisms of nonspherical hypersurfaces.

Parameters of automorphisms of hypersurfaces. Let two real analytic hypersurfaces M and M' be given by equations in normal form

$$v = \sum g_{\alpha\beta} z^\alpha \bar{z}^\beta + F(z, \bar{z}, u). \quad (1')$$

(F contains no terms of order lower than 4, no terms holomorphic or antiholomorphic in z , and satisfies some other conditions.)

Recall [3] that every locally biholomorphic mapping

$$f: M \rightarrow M', \quad f(0) = 0,$$

is in its lower-order terms equal to a linear fractional automorphism of the quadric:

$$z^* = \lambda_f U_f(z + a_f)/(1 - i\delta - r_f w), \quad w^* = \sigma_f \lambda_f^2 w/(1 - i\delta - r_f w), \quad (2)$$

where $\delta = 2 \sum g_{\alpha\beta} z^\alpha \bar{a}^\beta + \sum g_{\alpha\beta} a^\alpha \bar{a}^\beta$, $\sigma_f = \pm 1$, U_f is an $(n \times n)$ -matrix that transforms the form $\sum g_{\alpha\beta} z^\alpha \bar{z}^\beta$ into $\sigma_f \sum g_{\alpha\beta} z^\alpha \bar{z}^\beta$ (and so σ_f can be equal to -1 only if $n = 2p$), $\lambda_f > 0$ and r_f are real numbers, and $a_f = (a_f^1, \dots, a_f^n)$ —a vector in \mathbb{C}^n .

The higher-order terms of f are polynomials of the parameters $(U_f, \lambda_f, a_f, r_f)$ and the coefficients of the Taylor expansions of the equation of M , so one may try to investigate the properties of f by means of analysis of these relations. However, they seem very complicated, first of all, owing to the parameters a_f and r_f , which are in the denominator of the right-hand side of (2).

Since f is determined by the parameters $(U_f, \lambda_f, a_f, r_f)$, the dimension of the family of local biholomorphisms from M into M' and, in particular, the dimension of the local automorphism group of a hypersurface do not exceed

$(n^2 + 2n + 2)$. It turns out that this estimate is only sharp for quadrics. If the surfaces are nonspherical, there can be no more than n^2 independent real parameters and, what is important, the "bad" parameters a_f and r_f are among the dependent ones:

THEOREM 1 (BELOSHAPKA [4], LOBODA [5]). *If the real analytic surfaces M, M' are nonspherical, then the parameters λ_f, a_f, r_f are determined by the matrix U_f .*

Since $\lambda_f U_f$ is the restriction of the differential of f to the complex tangent hyperplane to M at the origin, one can restate the theorem as follows: a locally biholomorphic mapping of one nonspherical real analytic hypersurface into another is completely determined by the image of a point and by the conformal class of its differential's restriction to the complex tangent space at this point.

λ_f, a_f, r_f are in fact rational functions of the entries of U_f and the coefficients of the equations of M and M' . (The relations between these parameters are more specifically described in [6].) Let $M = M'$. The image of the group \mathcal{G} of local automorphisms of M under the map $f \mapsto U_f$ is a linear algebraic subgroup H isomorphic to \mathcal{G} . If $n \neq 2p$ then H is an algebraic subgroup of $U(p, n-p)$ —the group of matrices unitary with respect to the form $\sum g_{\alpha\beta} z^\alpha \bar{z}^\beta$. In particular, if $p = n$ then H is a compact group; hence \mathcal{G} is compact as well.

Using Theorem 1, one may estimate the main characteristics of the local biholomorphism f , e.g., its convergence radius and the maximum of its modulus. Unfortunately, as the proof of the theorem is based upon analysis of the normal forms of the equations, such estimates essentially depend on the order of zero of the nonspherical terms of M . In this way it is not even possible to provide uniform estimates for the points of M and M' close to the origin.

In the case of positive definite Levi form such estimates, however, really exist. They were obtained with the use of different methods that will be discussed in the following section.

Theorem on the germ of a mapping of pseudoconvex surfaces. Let M and M' be real analytic strictly pseudoconvex nonspherical hypersurfaces given by equations in normal form. As we mentioned in the preceding section, the set of all the local biholomorphisms from M into M' that fix the origin is compact. So there is a neighborhood of the origin in which all these mappings are defined and uniformly bounded. Certainly the size of the neighborhood and the bound on the norm depend on the size of the domains where the hypersurfaces are defined. Since the compactness is characteristic of nonspherical surfaces, it would be natural to suppose that these values depend also on how much these surfaces are nonspherical, that is, how large the nonspherical terms of their equations are in the vicinity of the origin. The result we are going to discuss states that these are the only data necessary to obtain the desired estimates.

We begin with definitions. For a hypersurface M given by an equation in normal form (1') with $F \not\equiv 0$, let $R(M)$ be the radius of the maximal ball in

\mathbb{C}^{2n+1} centered in the origin, where F (as a function of $(2n+1)$ independent variables) is defined and where its modulus does not exceed 1. For every positive $R \leq R(M)/2$ we will characterize the nonsphericity of M by the value $\chi(R, M)$ —the maximum of the modulus of F in the ball of radius R .

THEOREM 2 (A. G. VITUSHKIN [7]). *Let M, M' be strictly pseudoconvex nonspherical hypersurfaces given by equations in normal form. Then, for arbitrary $R \leq \min(R(M), R(M'))/2$ there are positive ε and m dependent only on $R, \chi(R, M), \chi(R, M')$ such that every locally biholomorphic $f: M \rightarrow M'$ extends to the ε -neighborhood of the origin and its norm in this neighborhood does not exceed m .*

To illustrate the theorem, consider two compact strictly pseudoconvex nonspherical hypersurfaces M and M' . Consider the family of mappings each defined on some open subset of M and mapping this subset biholomorphically into M' . The theorem states that this family is compact. Therefore, as an immediate consequence of Theorem 2, one obtains the theorem on the analytical continuation of locally defined maps of real analytic strictly pseudoconvex hypersurfaces (S. I. Pinchuk [8]; A. G. Vitushkin, V. V. Ezhov, N. G. Kruzhilin [9]).

The proof of the theorem essentially uses the properties of chains. These curves are CR-invariants of hypersurfaces, and they proved to be rather helpful in the study of biholomorphisms. In the real analytic case chains are closely related to the normal forms of the equations of surfaces, and this gives rise to distinguishing on each chain a family of parametrizations that we shall call linear normal. A different CR-invariant family of parametrizations on chains was introduced by Vitushkin. It is connected with another form of the equation of surfaces—the circular normal form (see [10])—so we shall call these parameters circular normal. Circular normal parameters, unlike the linear normal ones, extend along the whole chain and so they are more suitable for the problems of analytical continuation along the chains. We also note that there are ways of defining the normal parametrizations on chains also in the smooth case.

Through every point on a hypersurface in any direction transversal to the complex tangent space at this point, there exists a unique chain. The chains on hyperquadrics are their intersections with complex lines. In particular, the chain on a unit sphere which makes an angle α with the complex tangent at some point is a circle of radius $\sin \alpha$. The interval of the chain where the variation of any circular normal parameter equals 2π (or, equivalently, where some linear normal parameter varies from $-\infty$ to $+\infty$) corresponds to one circuit along this circle. We shall say that a chain γ on an arbitrary hypersurface makes one loop on a segment $\Delta\gamma$ if the variation of circular normal parameters on this segment is 2π . A CR-diffeomorphism takes such a segment to an analogous segment in the image.

Let M and M' be real analytic hypersurfaces, $x \in M, x' \in M'$, let the chain γ on M make one loop on a segment $\Delta\gamma$ that begins at x , and the chain $\gamma' \subset M'$ make one loop on a segment $\Delta\gamma'$ that begins at x' . It can be proved that every

locally biholomorphic mapping $f: M \rightarrow M'$ defined in no matter how small a neighborhood V of x and such that $f(x) = x'$, $f(\Delta\gamma \cap V) \subset \Delta\gamma'$, extends to the whole of $\Delta\gamma$ and maps it onto $\Delta\gamma'$.

On the other hand, if a chain through a point on a strictly pseudoconvex hypersurface makes a sufficiently small angle with the complex tangent space, then the hypersurface contains also a segment which begins at this point and on which the chain makes one loop. The reason is that such a segment is close to an analogous segment on a sphere, i.e., to a circle of a small radius. This can be stated more specifically (see [7, 12]); however it is sufficient for the proof of Theorem 2 that as the chain inclines to the complex tangent, the length of such a segment tends to zero.

For every chain γ through the origin, let us take the vector $(\eta_\gamma, 1)$ tangent to γ at the origin. It is easy to see that a biholomorphism f establishes between the chains on M and on M' the correspondence that has the form

$$\eta_{f(\gamma)} = \sigma_f \lambda_f^{-1} U_f(\eta_\gamma + a_f),$$

and so by means of this correspondence we may reconstruct the parameters λ_f, U_f, a_f . The correspondence between the normal parametrizations of M and M' is determined by the parameters λ_f and r_f . This observation together with the above properties of chains enables us to keep an eye on how much f stretches segments of chains, and that is the basis of the proof of the theorem.

Theorem 2 does not extend directly to hypersurfaces with indefinite Levi form. One reason is that in this case the differentials of local automorphisms taken in the fixed point may have arbitrarily large norms. It is not clear yet whether one could estimate the convergence radius and the maximum of the norm of a biholomorphism in terms of the nonsphericity characteristics of the surfaces and the norms of U_f and $(U_f)^{-1}$. Such an estimate might be useful in the situations related to analytic continuation of locally defined biholomorphisms of hypersurfaces with indefinite Levi form. Note that this continuation is in general impossible (Beloshapka [11]).

Linearization of local automorphisms of real analytic hypersurfaces. All local automorphisms of a quadric are of the form (2). If M is a nonspherical surface in normal form, the situation appears to be more complicated since, as it was mentioned above, the coefficients of local automorphisms are polynomials of their parameters and of the coefficients of the normal form. So the question whether there exists a coordinate system where all the automorphisms of M also have the form (2) may seem strange. However, it turns out that in many important cases the automorphisms can be reduced even to a linear form. We hope that in fact this is true for all nonspherical hypersurfaces.

The first case when the possibility of linearization was established was that of positive definite Levi form.

THEOREM 3 (KRUSHILIN, LOBODA [13]). *Let M be a real analytic strictly pseudoconvex nonspherical hypersurface. Then there exists a local coordinate system (z, w) where the equation of M is in normal form and every local automorphism has the form*

$$z^* = Uz, \quad w^* = w, \quad (3)$$

where U is a unitary matrix.

At present, we are able to obtain Theorem 3 as a straightforward corollary of two observations.

First, if the parameter a_f of a locally biholomorphic mapping $f: M \rightarrow M'$ of Levi nondegenerate hypersurfaces vanishes, then f is linear fractional. Suppose that f is a local automorphism. Then, as one can see easily, $a_f = 0$ and $\sigma_f = 1$ together with nonsphericity imply $r_f = 0$, so that f actually is of the form

$$z^* = \lambda_f U_f z, \quad w^* = \sigma_f \lambda_f^2 w, \quad (3')$$

and when $a_f = 0$ and $\sigma_f = -1$, then f can be reduced to the same form by a linear fractional coordinate change (see Ezhov [14]).

Let $a_f \neq 0$ and suppose that the automorphism f has a fixed chain γ through the origin. According to [3], there is a coordinate system where the equation of M has a normal form and γ coincides with the line $\{z = 0, v = 0\}$. In these coordinates, the parameter a_f vanishes. Therefore, f can be reduced to the form (3') if and only if it has a fixed chain.

Second, the compactness of the group \mathcal{G} of local automorphisms of strictly pseudoconvex nonspherical hypersurface M implies that, for $f \in \mathcal{G}$, $\lambda_f = 1$ and that the group of the affine transformations of \mathbb{C}^n :

$$\xi \mapsto \sigma_f \lambda_f^{-1} U_f (\xi + a_f)$$

also is compact and has a fixed point. Hence, the elements of \mathcal{G} have a common fixed chain.

In the coordinate system provided by Theorem 3, the group is realized as the group of all unitary symmetries of some set of real polynomials in \mathbb{C}^n . It is easy to see that the converse is also true: the group of unitary symmetries of an arbitrary set of real polynomials in \mathbb{C}^n is the local automorphisms group of some real analytic strictly pseudoconvex hypersurface in \mathbb{C}^{n+1} .

In the case of indefinite Levi form we do not have a final answer yet, but at least we know that each separate automorphism has a fixed chain, and so the following statement holds:

THEOREM 4 (EZHOV [14]). *Let M be a nonspherical hypersurface and f one of its local automorphisms. Then there exists a local coordinate system (z, w) where the equation of M is in normal form and f is of the form (3').*

Suppose the matrix U_f is equivalent to a diagonal one. Then the assumption that there are no fixed chains so that the equation

$$\sigma_f \lambda_f^{-1} U_f (\xi + a_f) = \xi$$

is insolvable leads to a contradiction with the algebraic dependence of a_{f^k} on the entries of $(U_f)^k$. The case when the Jordan normal form of U_f is not diagonal is treated with the use of a technique similar to that in the proof of Theorem 1. Besides that, one has to consider the problem of reducing simultaneously an indefinite Hermitian form in \mathbb{C}^n and an operator unitary with respect to it to suitable forms.

Besides strictly pseudoconvex hypersurfaces, the existence of a fixed chain common to all automorphisms was proved for nonspherical hypersurfaces of signature $(n-1, 1)$ (Ezhov [15]). The proof is rather complicated: it involves detailed analysis of transformations of normal forms under coordinate changes and some results of the classical theory of invariants. In the corresponding coordinate system all the automorphisms have the form (3') and \mathcal{G} is realized as the group of all the linear transformations of the form λU (where $\lambda > 0$ and $U \in U(n-1, 1)$) that are symmetries of some set of real homogeneous rational functions on \mathbb{C}^n .

Notice that for some subclass the class of nonspherical hypersurfaces D. Burns, S. Shnider, and R. Wells [16] and S. Webster [17] proved the existence of local coordinate systems where the equations of these surfaces are in normal form and satisfy some additional restrictions. The changes of coordinates between these coordinate systems have the form (3) with $U \in U(p, n-p)$, so every local automorphism and every biholomorphic mapping of such a hypersurface into another one have the same form (the parameters λ_f and σ_f of automorphisms of these surfaces are always equal to 1). However, it is not clear yet whether there are similar special normal forms for all nonspherical hypersurfaces.

Automorphisms of smooth hypersurfaces. In the above sections we dealt with real analytic hypersurfaces. Let us now turn to smooth ones.

Suppose that M and M' are two smooth hypersurfaces in \mathbb{C}^{n+1} with non-degenerate Levi forms, and let $f: M \rightarrow M'$ be a smooth locally defined CR-diffeomorphism. Choose coordinate systems where the equations of M and M' are in normal form up to the terms of a high order. In these coordinates, as well as in the real analytic case, f coincides in its lower terms with a linear fractional map (2) and is completely determined by the parameters

$$(U_f, \lambda_f, a_f, r_f).$$

Using the Taylor series of the equations of M and M' , one can construct their formal normal forms. If these forms have nonspherical terms, then Theorem 1 extends immediately to this situation. However, a C^k -regular hypersurface can be nonspherical in a point and have the order of contact with a quadric equal to k . In this case, one cannot apply the techniques based on the consideration of normal forms. Nevertheless, the following holds:

THEOREM 5 [12]. *If f is a smooth local CR-automorphism of a smooth strictly pseudoconvex nonspherical hypersurface M , then $\lambda_f = 1$ and the parameters a_f and r_f are determined by the matrix U_f .*

The main tools of the proof are chains and integrals over chains of an invariant real 1-form introduced by Webster in [18, 19]. The nonsphericity of M implies that in no neighborhood of the distinguished point is this form θ identically equal to zero. The integral of θ over any segment of a chain $\gamma \subset M$ (oriented in accordance with normal parametrizations) that intersects the domain where $\theta \neq 0$ is strictly positive. If such a segment $\Delta\gamma$ lies in the domain of f , then

$$\int_{f(\Delta\gamma)} \theta = \int_{\Delta\gamma} \theta$$

so the length of $f(\Delta\gamma)$ cannot be too small.

To apply this observation, it is necessary to find a segment of a chain connecting the distinguished point with some point in which $\theta \neq 0$. It turns out to be always possible on strictly pseudoconvex surfaces:

THEOREM 6 [12]. *Let M be a strictly pseudoconvex hypersurface and let x be a point of M . For every $\varepsilon > 0$ the set of points that can be connected with x by a segment of a chain making less than one loop and of the length less than ε contains a neighborhood of the point x .*

The fact that any two sufficiently close points on a pseudoconvex surface can be joined by a chain was also established by H. Jacobowitz [20].

The proof of Theorem 6 is based upon the closeness of the chains on the hypersurface and on the sphere. The required estimates (more precise than those used in the proof of Theorem 2) were obtained by means of analysis of the system of differential equations for chains proposed by Burns and Shnider [21].

Since the result of Theorem 1 holds on smooth pseudoconvex surfaces, one may try to extend to this case also the results on the existence of fixed chains and linearization of automorphisms.

THEOREM 7 [12]. *Let M be a smooth strictly pseudoconvex nonspherical hypersurface. Then there exists a chain through the distinguished point which is fixed under all local automorphisms of M .*

Note that on a smooth hypersurface the group \mathcal{G} is not compact in general, since there may be no neighborhood of the fixed point where all the automorphisms are defined. One has to use other ideas. First, the same methods as in the proof of Theorem 6 apply to show that every local automorphism of M has a fixed chain. This implies that every automorphism generates a compact subgroup of \mathcal{G} ; therefore the group \mathcal{G}^0 —the identity component of \mathcal{G} —is compact, and there are chains on M fixed under automorphisms from \mathcal{G}^0 . Further considerations use the theorem of I. Schur on linear periodic groups.

COROLLARY TO THEOREM 7. *For every neighborhood V of the distinguished point on M there is a neighborhood $W \subset V$ such that every local automorphism of M defined in W maps W onto itself. These automorphisms form a compact subgroup of \mathcal{G} . In some local coordinates $(z = (z^1, \dots, z^n), u)$ on M , every*

transformation f from this family is of the form

$$(z, u) \mapsto (U_f z, u)$$

where U_f is a unitary matrix.

The above results relate to smooth strictly pseudoconvex hypersurfaces. Examples show [12] that even the statement of Theorem 1 is not valid on smooth hypersurfaces with indefinite Levi form. On the whole, we would say that the indefiniteness of the Levi form adds difficulties to the investigations of the local biholomorphisms. In some cases, however, this indefiniteness implies interesting properties of biholomorphisms. As an example, we will state a result, of global rather than of local nature, but that fits in well with our discussion of linearizations of locally biholomorphic maps:

THEOREM 8 (VITUSHKIN, IVASHCOVITCH [22]). *Let M and M' be compact real hypersurfaces in \mathbb{CP}^{n+1} and let their Levi forms have eigenvalues of opposite signs in each point. Then every holomorphic mapping $f: M \rightarrow M'$ which is locally biholomorphic in the vicinity of each point of M extends to a linear automorphism of \mathbb{CP}^{n+1} .*

REFERENCES

1. Henri Poincaré, *Les fonctions analytiques de deux variables et la représentation conforme*, Rend. Circ. Mat. Palermo **23** (1907), 185–220; reprinted in his *Oeuvres*, Vol. IV, Gauthier-Villars, Paris, 1950, pp. 244–289.
- 2a. Élie Cartan, *Sur la géométrie pseudo-conforme des hypersurfaces de l'espace de deux variables complexes*. I, Ann. Mat. Pura Appl. (4) **11** (1932/33), 17–90; reprinted in his *Oeuvres complètes*, Part II, Vol. 2, Gauthier-Villars, Paris, 1953 (reprint, Éditions du CNRS, Paris, 1984), pp. 1231–1304.
- 2b. —, *Sur la géométrie pseudo-conforme des hypersurfaces de l'espace de deux variables complexes*. II, Ann. Scuola Norm. Sup. Pisa (2) **1** (1932), 333–354; reprinted in his *Oeuvres complètes*, Part III, Vol. 2, Gauthier-Villars, Paris, 1955 (reprint, Éditions du CNRS, Paris, 1984), pp. 1217–1238.
3. S. S. Chern and J. K. Moser, *Real hypersurfaces in complex manifolds*, Acta Math. **133** (1974/75), 219–271.
4. V. K. Beloshapka, *The dimension of the group of automorphisms of an analytic hypersurface*, Izv. Akad. Nauk SSSR Ser. Mat. **43** (1979), 243–266=Math. USSR Izv. **14** (1980), 223–245.
5. A. V. Loboda, *On local automorphisms of real analytic hypersurfaces*, Izv. Akad. Nauk SSSR Ser. Mat. **45** (1981), 620–645=Math. USSR Izv. **18** (1982).
6. V. K. Beloshapka and A. G. Vitushkin, *Estimates for the radius of convergence of power series defining mappings of analytic hypersurfaces*, Izv. Akad. Nauk SSSR Ser. Mat. **45** (1981), 962–984=Math. USSR Izv. **19** (1982), 241–259.
7. A. G. Vitushkin, *Real-analytic hypersurfaces of complex manifolds*, Uspekhi Mat. Nauk **40** (1985), no. 2 (242), 3–31=Russian Math. Surveys **40** (1985), no. 2, 1–35.
8. S. I. Pinchuk, *On holomorphic mappings of real analytic hypersurfaces*, Mat. Sb. **105**(147) (1978), 574–593=Math. USSR Sb. **34** (1978), 503–519.
9. A. G. Vitushkin, V. V. Ezhov, and N. G. Kruzhilin, *Extension of local mappings of pseudoconvex surfaces*, Dokl. Akad. Nauk SSSR **270** (1983), 271–274=Soviet Math. Dokl. **27** (1983), 580–583.
10. A. G. Vitushkin, *Global normalization of a real-analytic surface along a chain*, Dokl. Akad. Nauk SSSR **269** (1983), 15–18=Soviet Math. Dokl. **27** (1983), 270–273.

11. V. K. Beloshapka, *Example of a noncontinuable holomorphic transformation of an analytic hypersurface*, Mzt. Zametki **32** (1982), 121–123=Math. Notes **32** (1982), 540–541.
12. N. G. Kruzhilin, *Local automorphisms and mappings of smooth strictly pseudoconvex hypersurfaces*, Izv. Akad. Nauk SSSR Ser. Mat. **49** (1985), 566–591=Math. USSR Izv. **26** (1986), 531–552.
13. N. G. Kruzhilin and A. V. Loboda, *Linearization of local automorphisms of pseudoconvex surfaces*, Dokl. Akad. Nauk SSSR **271** (1983), 280–282=Soviet Math. Dokl. **28** (1983), 70–72.
14. V. V. Ezhov, *On the linearization of automorphisms of a real analytic hypersurface*, Izv. Akad. Nauk SSSR Ser. Mat. **49** (1985), 731–765=Math. USSR Izv. **27** (1986), 53–84.
15. ———, *Linearization of the stability group of a class of hypersurfaces*, Uspekhi Mat. Nauk **41** (1986), no. 3 (249), 181–182=Russian Math. Surveys **41** (1986), no. 3 (to appear).
16. D. Burns, S. Shnider, and R. O. Wells, *Deformations of strictly pseudoconvex domains*, Invent. Math. **46** (1978), 237–253.
17. S. M. Webster, *On the Moser normal form at a non-umbilic point*, Math. Ann. **233** (1978), 97–102.
18. ———, *On the transformation group of a real hypersurface*, Trans. Amer. Math. Soc. **231** (1977), 179–190.
19. ———, *Pseudo-Hermitian structures on a real hypersurface*, J. Differential Geom. **13** (1978), 25–41.
20. H. Jacobowitz, *Chains in CR geometry*, J. Differential Geom. **21** (1985), 163–195.
21. D. Burns and S. Shnider, *Real hypersurfaces on complex manifolds*, Proc. Sympos. Pure Math., vol. 30, part 2, Amer. Math. Soc., Providence, R.I., 1977, pp. 141–168.
22. A. G. Vitushkin and S. M. Ivashkovich, *On the extension of holomorphic mappings of a real analytic hypersurface into complex projective space*, Dokl. Akad. Nauk SSSR **267** (1982), 779–780=Soviet Math. Dokl. **26** (1982).

STEKLOV MATHEMATICS INSTITUTE, MOSCOW, USSR

Complex Geometry in Convex Domains

LÁSZLÓ LEMPert

1. First we shall briefly discuss real (projective) geometry in convex domains. Let D be a bounded convex domain in real Euclidean space \mathbf{R}^n . Hilbert has introduced a metric on D in the following way. Let x and y be points in D and let the straight line through x and y intersect the boundary ∂D in p and q . Put

$$h(x, y) = |\log(x y p q)| = \left| \log \left(\frac{xp}{xq} : \frac{yp}{yq} \right) \right|.$$

This is the Hilbert distance of x and y . For instance when D is a ball, the metric space (D, h) is a model of hyperbolic geometry, the Cayley-Klein model, in which case h comes from a Riemann metric. In general h is just a Finsler metric. Its basic property is that it is invariant under projective transformations: If $\phi: D \rightarrow D'$ is a one-to-one projective transformation then it is an isometry between the Hilbert metrics of the domains D, D' .

S. Kobayashi in [15] proposed other ways to define the same metric. First a gauge will be introduced to measure distances with. Our gauge is the open interval $I = (-1, 1) \subset \mathbf{R}$. The distance between two points $\xi, \eta \in I$ is defined to be

$$d(\xi, \eta) = |\log(\xi, \eta, -1, 1)| = \left| \log \left(\frac{\xi + 1}{\xi - 1} : \frac{\eta + 1}{\eta - 1} \right) \right|.$$

This is of course just Hilbert's metric for $D = I$. It is invariant under projective automorphisms of I ; more generally if ϕ is a projective mapping of I into itself, then $d(\phi(\xi), \phi(\eta)) \leq d(\xi, \eta)$.

To measure the distance between two points x and y in an arbitrary bounded, convex domain $D \subset \mathbf{R}^n$, place the gauge in D so that it goes through x and y and read the distance on the gauge between x and y . Here by "placing" we mean a projective mapping of I into D . Of course, there are several ways to map I into D ; accordingly we shall record different readings. $\bar{h}(x, y)$ is defined to be the smallest of all the readings:

$$\begin{aligned} \bar{h}(x, y) = \inf \{ & d(\xi, \eta) \mid \text{there is a projective } f: I \rightarrow D \\ & \text{s.t. } f(\xi) = x, f(\eta) = y \}. \end{aligned} \quad (1)$$

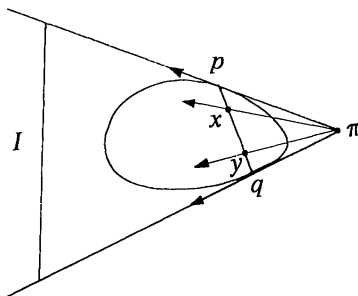


FIGURE 1

It is easy to see that in (1) the extremum is attained and the extremal maps f map I on the intersection of D with the straight line through x and y . This also shows that $\bar{h}(x, y)$ and $h(x, y)$ agree.

A dual way of measuring distances in D is to map D into I rather than I into D , in a projective way, and to ask how far x and y can go; i.e., put

$$\underline{h}(x, y) = \sup\{d(F(x), F(y)) \mid F: D \rightarrow I \text{ is projective}\}. \quad (2)$$

It is quite straightforward that $\underline{h}(x, y) \leq \bar{h}(x, y)$. Indeed, if f and F are mappings as in (1), resp. (2), then $F \circ f$ is a projective mapping of I into itself so that by the contracting property of the distance d

$$d(F(x), F(y)) = d(F(f(\xi)), F(f(\eta))) \leq d(\xi, \eta).$$

Taking the sup of the left-hand side and the inf of the right-hand side $\underline{h} \leq \bar{h}$ is obtained.

Till now we have not used convexity. It turns out that convexity implies that $\underline{h}(x, y)$ and $\bar{h}(x, y)$ agree. Indeed, let p and q denote, as before, the points of intersection of D with the line through x and y . Then, as observed before, pq is the extremal position of the gauge, so that $\bar{h}(x, y) = |\log(xypq)|$. Let the tangent hyperplanes to ∂D in p and q intersect in the $(n-2)$ plane π (see Figure 1). Place I between the two tangent hyperplanes so that its endpoints lie on these hyperplanes, and project D onto I along hyperplanes that contain π . With this projection F

$$\underline{h}(x, y) \geq d(F(x), F(y)) = |\log(xypq)| = \bar{h}(x, y),$$

so that $\underline{h} = \bar{h}$ as stated.

For later reference we note that a crucial point of this argument was that the tangent hyperplanes in p and q could be included into a projective family of hyperplanes (those passing through π). This is of course a trivial fact in the present context since *any two* hyperplanes determine a projective pencil of hyperplanes.

2. The complex geometry referred to in the title is obtained by replacing projective concepts by holomorphic ones in the preceding discussion. Thus our gauge now will be the unit disc $U \subset \mathbb{C}$ endowed with the hyperbolic distance

$$\text{hyp}(\zeta, \omega) = |\log(\zeta, \omega, \pi, \rho)|,$$

π and ρ denoting the points where the circle ∂U intersects that circle through $\zeta, \omega \in U$ which is orthogonal to ∂U . Thus (U, hyp) is the Poincaré model of the hyperbolic plane. Using this gauge C. Carathéodory in the twenties and S. Kobayashi in the sixties defined complex analogues of Hilbert's metric—although the realization of the analogy came apparently later than the definitions themselves (see [15]).

Restricting ourselves to smoothly (C^∞) bounded strictly convex domains $D \subset \mathbb{C}^n$, the Kobayashi distance of $z, w \in D$ ($z \neq w$) is defined to be

$$k(z, w) = \inf \{ \text{hyp}(\zeta, \omega) \mid \text{there is a holomorphic } f: U \rightarrow D \\ \text{s.t. } f(\xi) = z, f(\omega) = w \} \quad (3)$$

(see [6, 7]). Dually, the Carathéodory distance is defined by

$$c(z, w) = \sup \{ \text{hyp}(F(z), F(w)) \mid F: D \rightarrow U \text{ is holomorphic} \} \quad (4)$$

(see [3]). Again, it is not hard to see that the extrema in (3) and (4) are attained. That these geometries can be investigated depends on the fact that it is possible to characterize the extremal mappings f in (3). Indeed, the following theorem holds:

THEOREM 1. *A holomorphic mapping $f: U \rightarrow D$ is extremal (of the variational problem (3)) if and only if it satisfies the following three conditions.*

1. *f smoothly extends to the closed disc \bar{U} ;*
2. *$f(\partial U) \subset \partial D$;*
3. *The family $T_{f(\zeta)}^{\mathbb{C}}(\partial D)$ ($\zeta \in \partial U$) of complex tangent hyperplanes to ∂D in $f(\zeta)$ can be included into a smooth, holomorphic family $\{E_{f(\zeta)}: \zeta \in \bar{U}\}$ of hyperplanes.*

Condition 3 means more precisely that there is a smooth family $\{E_{f(\zeta)}: \zeta \in \bar{U}\}$ of hyperplanes in \mathbb{C}^n such that $E_{f(\zeta)}$ depends holomorphically on $\zeta \in U$ and, for $\zeta \in \partial U$, $E_{f(\zeta)}$ is tangent to ∂D in $f(\zeta)$ (and therefore coincides with $T_{f(\zeta)}^{\mathbb{C}}(\partial D)$).

It is easy to see that in this case $f(\zeta) \in E_{f(\zeta)}$ for all $\zeta \in \bar{U}$, so that $E_{f(\zeta)}$ are the fibres of a smooth holomorphic subbundle E of the restricted bundle $T\mathbb{C}^n|_{f(\bar{U})}$, where $T\mathbb{C}^n$ is the holomorphic tangent bundle of \mathbb{C}^n .

The proof of this theorem runs as follows. First it is computed that it is true in the case when D is a ball. Then it is proved that if it is true for a domain D_0 then it is also true for all domains that can be obtained from D_0 by a continuous deformation (through smoothly bounded strictly convex domains); see [7, 9]. A more direct approach has been proposed by H. Royden and P. Wong in [13]. E. Poletski considers the more general problem when D is strictly pseudoconvex. In this context Theorem 1 is completely false but weaker results are available (see [12]).

One corollary of Theorem 1 is that Carathéodory's and Kobayashi's metrics agree. Indeed, $k(z, w) \geq c(z, w)$ follows by Schwarz's lemma, just as in §1. The

converse inequality also follows along the same lines as there, the fibres of the extremal projection $F: D \rightarrow U$ being this time the hyperplanes $E_{f(\zeta)}$ (see [8, 9, 13]).

3. If $f: U \rightarrow D$ is extremal, call $f(U)$ an extremal disc. A more or less direct consequence of Theorem 1 is that through any couple of points $z, w \in D$ ($z \neq w$) there goes a unique extremal disc. The corresponding extremal mapping is unique up to composition by a holomorphic automorphism of U . Extremal maps smoothly embed \overline{U} into \overline{D} , and these embeddings (restricted to U) are isometric between the hyperbolic metric of U and the Kobayashi (or Carathéodory) metric of D . Finally, given a point $z \in D$ and a direction $v \in T_z(D)$ in z , $v \neq 0$, there is a unique extremal disc that goes through z and is tangent to v . This extremal disc varies smoothly when z and v vary smoothly. For all this the reference is [7, 9].

4. Many other properties of the above complex geometry can be deduced from Theorem 1 (or its proof). Rather than listing them we shall mention three instances where notions connected with this geometry can be applied with success. We are only going to touch upon two of them, while the third will be discussed at some length.

5. Consider the boundaries of extremal discs in D , i.e., the intersection of all closed extremal discs $f(\overline{U})$ with ∂D . This gives a network of smooth Jordan curves on ∂D , a holomorphically invariant structure on ∂D . Smoothness properties of this structure can be used to study boundary regularity of biholomorphic mappings, in particular to prove precise boundary regularity for biholomorphic mappings between strictly pseudoconvex domains whose boundaries satisfy a finite differentiability condition— a quantitative version of C. Fefferman's theorem in [4] (see [10], but also S. V. Hasenoff's paper [5], the details of which we, however, ignore).

6. For Z a fixed point in D , consider all extremal mappings $f: \overline{U} \rightarrow \overline{D}$ such that $f(0) = Z$. Along these mappings push forward the Green function of \overline{U} with pole in 0 (i.e., $\log |\zeta|$) to get a smooth function $u: \overline{D} \rightarrow \mathbf{R}$. This function can be considered as the Green function of D for the nonlinear potential theory associated with the Monge-Ampère operator $\det \partial^2 / \partial z_j \partial \bar{z}_k$, as proposed in [2] by E. Bedford and B. A. Taylor. Indeed, u satisfies

$$\begin{aligned} u &\text{ is plurisubharmonic in } D, \\ \det \partial^2 u / \partial z_j \partial \bar{z}_k &= 0 \quad \text{in } D \setminus \{Z\}, \\ u &= 0 \quad \text{on } \partial D, \\ u(z) &= \log |z - Z| + O(1) \quad \text{as } z \rightarrow Z. \end{aligned}$$

7. The last problem we shall deal with is the following: How many holomorphically inequivalent n -dimensional (convex) domains are there? When $n = 1$, by virtue of the Riemann mapping theorem there is just one (if we restrict ourselves to bounded domains), whereas when $n \geq 2$ there are infinitely many (see [14] or for a more precise result, [1]). It is of course not really the cardinality of the set of equivalence classes that one is interested in. The problem is much rather to impose some natural structure on this set and investigate properties of the structure defined.

Not surprisingly, it will be easier to consider marked domains rather than plain domains. A marked domain is a domain D with a base point $p \in D$ and an n -frame of $T_p(D)$. Two marked domains are equivalent if there is a biholomorphism between the two under which the base points and the n -frames in them correspond. For one thing, introducing marking has the advantage that it eliminates nontrivial automorphisms of domains, a constant nuisance when constructing moduli spaces.

Equivalently, one could consider strictly convex, smoothly bounded domains $D \subset \mathbb{C}^n$ that contain the origin and call two such domains equivalent if there is a biholomorphism ϕ between the two such that $\phi(0) = 0$ and the differential $\phi_*(0)$ is the identity. Denote the set of such domains by \mathfrak{X} and the corresponding equivalence relation by \sim . \mathfrak{X} carries a natural topology, that of the C^∞ -convergence of domains. Understanding the factor topology \mathfrak{X}/\sim can be considered as an n -dimensional variant of the Riemann mapping theorem.

Using properties of extremal mappings and invariants associated with them one can prove the following theorems.

THEOREM 2. *For arbitrary n , \mathfrak{X}/\sim is homeomorphic to a contractible subset of a Fréchet space.*

THEOREM 3. *When $n = 2$, \mathfrak{X}/\sim is homeomorphic to an open neighborhood of the zero section in a smooth Fréchet bundle $\mathcal{E} \rightarrow \mathcal{C}$ whose base \mathcal{C} is an open convex cone in some Fréchet space.*

Almost nothing is known about the size of the neighborhood. It might be the entire total space \mathcal{E} . The picture is this. With *every* point $e \in \mathcal{E}$ one can associate a two-dimensional (marked) Stein manifold $M(e)$. Some of these manifolds, and maybe all, admit a convex embedding into \mathbb{C}^2 . On the other hand, an arbitrary strictly convex marked domain in \mathbb{C}^2 , with smooth boundary, is equivalent to one and only one manifold $M(e)$. For example, circular domains—those that are left invariant by rotations $z \mapsto e^{it}z$ —are in a one-to-one correspondence with points $e \in \mathcal{E}$ on the zero section.

8. We shall now indicate the proof of Theorem 3. Let $D \in \mathfrak{X}$. For $r > 0$ put

$$D_r = \{z \in D: k(0, z) < r\}, \quad \frac{1}{r}D_r = \{z/r: z \in D_r\}.$$

As r goes to 0, the domains $\frac{1}{r}D_r$ converge in \mathfrak{X} to a circular domain $J \in \mathfrak{X}$, called the indicatrix of D (in 0). One can think of the indicatrix as the best circular approximation of D .

The indicatrix is an invariant of D ; i.e., if $D \sim D'$ then the corresponding indicatrices J and J' coincide. Denote the subset of \mathfrak{X} consisting of all circular domains by \mathcal{C} . This is a topological cone with respect to the usual addition of convex domains and, in fact, can be realized as an open convex cone in some Fréchet space \mathcal{F} . A slightly different realization of \mathcal{C} as an open subset in some Fréchet space is due to G. Patrizio and P. Wong (see [11]).

We shall next define another invariant of D , which is going to measure the deviation of D from being circular. This invariant will be a quadratic form q on a line bundle $T \rightarrow \partial J$, whose fibre T_v above $v \in \partial J$ consists of $(1, 0)$ vectors $c_1 \partial/\partial z_1 + c_2 \partial/\partial z_2$ tangential to ∂J in v . Rather than defining q on each individual fibre T_v , we shall define q simultaneously on the fibres $T_{\zeta v}$, $|\zeta| = 1$, where $v \in \partial J$ is fixed.

In order to do so, take a $v \in \partial J$ and suppose first D is of a very special form. Namely, suppose that one can introduce linear coordinates w_1, w_2 in \mathbb{C}^2 in such a way that a suitable defining function r of D is of the form

$$r(w) = -1 + |w_1|^2 + |w_2|^2 + \operatorname{Re} a(w_1)w_2^2 + O(w_2^3), \quad (5)$$

with a a smooth function of w_1 , and the coordinates of v are $(1, 0)$ (in the system w_1, w_2). In this case Theorem 1 yields that the mapping $f(\zeta) = (\zeta, 0)$ is an extremal mapping, and we can also compute extremal mappings infinitesimally near to f . Hence one has a description of the indicatrix J in an infinitesimal neighborhood of the circle $\{\zeta v: |\zeta| = 1\} \subset \partial J$. It turns out that T_v , and more generally $T_{\zeta v}$, $|\zeta| = 1$, consists of vectors parallel to $\partial/\partial w_2$.

Returning to (5), note that r defines a circular domain if $a(w_1) = \bar{w}_1^2$. Since it is really the restriction of $a(w_1)$ to $|w_1| = 1$ that matters, this suggests to expand $a(w_1)$ into a Fourier series

$$a(w_1) = \sum_{-\infty}^{\infty} a_j w_1^j \quad (|w_1| = 1)$$

and consider the coefficients a_j ($j \neq -2$) as the deviation of D from being circular. It turns out that of these coefficients only those with $j < -2$ are holomorphically invariant. Therefore we form

$$A(w_1) = \sum_{-\infty}^{-3} a_j w_1^j \quad (6)$$

and put

$$q(\zeta v)(c_2 \partial/\partial w_2) = A(\zeta) c_2^2 \quad (|\zeta| = 1). \quad (7)$$

This is the definition of the quadratic form $q(\zeta v)$ on $T_{\zeta v}$ ($|\zeta| = 1$), when D is defined by a function r of the special form (5).

When $D \in \mathfrak{X}$ is arbitrary, one can always find a domain $D' \in \mathfrak{X}$ equivalent to it and of the form treated above; if $q'(\zeta v)$ is the quadratic form for D' , just put $q(\zeta v) = q'(\zeta v)$. It can be checked that this definition of $q(\zeta v)$ does not depend on which particular D' was picked.

Thus we have associated with every $D \in \mathfrak{X}$ a pair (J, q) , where $J \in \mathcal{C}$ and q is a smooth quadratic form on a certain line bundle $T \rightarrow \partial J$. This quadratic form has the special feature that for any $v \in \partial J$ the function (or more precisely, section)

$$\partial U \ni \zeta \mapsto q(\zeta v) \quad (8)$$

has an antiholomorphic extension to U , which vanishes in the origin to third order (see (6) and (7)).

Consider now for an arbitrary $J \in \mathcal{C}$ the Fréchet space of those smooth quadratic forms q on $T \rightarrow \partial J$ that have the above property; i.e., for any $v \in \partial J$ the function (8) has an antiholomorphic extension to U , vanishing in the origin to third order. Call this space \mathcal{E}_J . On the union $\bigcup \{\mathcal{E}_J : J \in \mathcal{C}\}$ there is a natural structure of a Fréchet bundle $\mathcal{E} \rightarrow \mathcal{C}$. By what has been said above, we have a mapping $j: \mathfrak{X}/\sim \rightarrow \mathcal{E}$ which sends an equivalence class $[D] \in \mathfrak{X}/\sim$ to the pair (J, q) consisting of the indicatrix J of D and the corresponding quadratic form q on the bundle $T \rightarrow \partial J$.

This is the construction underlying Theorem 3. The proof is then concluded by verifying that the mapping j is a homeomorphism onto an open neighborhood of the zero section of \mathcal{E} . Details of this proof will appear elsewhere.

REFERENCES

1. D. Burns, S. Shnider, and R. O. Wells, *On deformations of strictly pseudoconvex domains*, Invent. Math. **46** (1978), 237–253.
2. E. Bedford and B. A. Taylor, *The Dirichlet problem for the complex Monge-Ampère equation*, Invent. Math. **37** (1976), 1–44.
3. C. Carathéodory, *Über das Schwarzsche Lemma bei analytischen Funktionen von zwei komplexen Veränderlichen*, Math. Ann. **97** (1926), 76–98.
4. C. Fefferman, *The Bergman kernel and biholomorphic mappings of pseudoconvex domains*, Invent. Math. **26** (1974), 1–65.
5. S. V. Hasanoff (to appear).
6. S. Kobayashi, *Hyperbolic manifolds and holomorphic mappings*, Marcel Dekker, New York, 1970.
7. L. Lempert, *La métrique de Kobayashi et la représentation des domaines sur la boule*, Bull. Soc. Math. France **109** (1981), 427–474.
8. —, *Holomorphic retracts and intrinsic metrics in convex domains*, Anal. Math. **8** (1982), 257–261.
9. —, *Intrinsic distances and holomorphic retracts*, Complex Analysis and Applications '81 (Sofia, 1981), 1984.
10. —, *A precise result on the boundary regularity of biholomorphic mappings*, Math. Z. **193** (1986), 559–579.
11. G. Patrizio and P. Wong, *Stability of the Monge-Ampère foliation*, Math. Ann. **263** (1983), 13–29.
12. E. A. Poletski, *The Euler-Lagrange equations for the extremal holomorphic mappings of the unit disc*, Michigan Math. J. **30** (1983), 317–333.
13. H. Royden and P. Wong, *Carathéodory and Kobayashi metric on convex domains*, Preprint.
14. P. Thullen, *Zu den Abbildungen durch analytische Funktionen mehrerer Veränderlichen*, Math. Ann. **104** (1931), 244–259.
15. S. Kobayashi, *Intrinsic distances associated with flat affine or projective structures*, J. Fac. Sci. Univ. Tokyo **24** (1977), 120–135.

Metric Properties of Harmonic Measure

N. G. MAKAROV

Let Ω be a planar domain whose boundary has positive capacity, and let ω be the harmonic measure on $\partial\Omega$. We will discuss metric properties of sets supporting ω . There has been considerable interest in this topic for a long time and many people contributed to its development. We shall report on some recent advances. The main progress is based upon: (i) the relevance, for the simply-connected domains, of certain results on the derivative of a univalent function; (ii) the idea to use critical points of Green's function when Ω is very nonsimply-connected; (iii) the application of dynamical systems' methods in scale invariant cases. It should be pointed out that the research, on the whole, was greatly influenced by Carleson's ideas presented in [4] and [5].

We are primarily concerned with a question of how ω is connected with Hausdorff measures Λ_ϕ corresponding to various measure functions ϕ . We say that ω is absolutely continuous with respect to Λ_ϕ and write $\omega \ll \Lambda_\phi$ if $\Lambda_\phi(e) = 0$ implies $\omega(e) = 0$. On the other hand, ω is said to be singular with respect to Λ_ϕ (notation: $\omega \perp \Lambda_\phi$) if there exists a set of Λ_ϕ -measure zero supporting ω . By definition, $\dim \omega = \inf\{p \geq 0: \omega \perp \Lambda_p\}$, where we write Λ_p instead of Λ_ϕ if $\phi(t) = t^p$.

In the first three sections of this paper we deal with simply-connected domains. The fourth section is devoted to the general case; we are rather brief in describing the outstanding results of Jones and Wolff [10] since we may refer to the authors' reports. In the final section we discuss a dynamics counterpart of the theory considered.

1. Simply-connected case: connections with the derivative of conformal mapping. Let f be a univalent function mapping the unit disc \mathbf{D} into a simply-connected domain Ω . Then $f(\zeta)$, $\zeta \in \partial\mathbf{D}$, exists as an angular limit a.e. on $\partial\mathbf{D}$, and for $e \subset \partial\Omega$ we define $f^{-1}e = \{\zeta \in \partial\mathbf{D}: f(\zeta) \text{ exists and } \in e\}$. The harmonic measure ω evaluated at $f(0)$ is given by $\omega(e) = |f^{-1}e|$. Thus metric properties of ω depend on the boundary distortion of f , whereas the latter depends on the behavior of the derivative. To study this relation between ω and f' more closely, we first consider a simpler case when Ω is a quasidisc, i.e., when

$\partial\Omega$ is a Jordan curve such that for any z_1 and z_2 on $\partial\Omega$, the diameter of the smaller arc between z_1 and z_2 does not exceed $\text{const}|z_1 - z_2|$. Then for any arc I of $\partial\mathbf{D}$ we have $\text{diam } fI \asymp |I||f'(a_I)|$, where $a_I = (1 - |I|)\zeta_I$ and ζ_I is the center of I . Hence $\Lambda_\phi(e) = 0$ iff the set $E = f^{-1}e$ satisfies

$$\inf \left\{ \sum \phi(|I||f'(a_I)|) : E \subset \bigcup I \right\} = 0, \quad (1)$$

the infimum being taken over all coverings of E . Suppose, for simplicity, that $\phi(2t) \leq \text{const } \phi(t)$ and $\phi(\varepsilon t) = o(\phi(t))$ as $\varepsilon \rightarrow 0$, and let $\psi = \psi(t)$ be defined by $\phi(t\psi(t)) \equiv t$. If $|f'(r\zeta)| \geq \text{const } \psi(1-r)$ on E , then

$$\sum \phi(|I||f'(a_I)|) \geq \text{const } \sum |I|$$

and hence (1) implies $\omega(e) = 0$. On the other hand, if $\liminf_{r \rightarrow 1} |f'(r\zeta)|/\psi(1-r) = 0$ on a set $E \subset \partial\mathbf{D}$, then, by the covering lemma, E satisfies (1). In other words,

$$\begin{aligned} \omega \ll \Lambda_\phi &\Leftrightarrow |f'(r\zeta)| \geq c_\zeta \psi(1-r) \quad \text{a.e. on } \partial\mathbf{D}; \\ \omega \perp \Lambda_\phi &\Leftrightarrow \liminf_{r \rightarrow 1} |f'(r\zeta)|/\psi(1-r) = 0 \quad \text{a.e. on } \partial\mathbf{D}. \end{aligned} \quad (2)$$

In the general case we have only the following partial result.

THEOREM 1. *If ϕ is a logarithmico-exponential function, then the equivalences (2) hold for any simply-connected domain.*

The hypothesis certainly may be weakened considerably but it is not clear whether the assertion remains valid without any restriction on ϕ . The proof in [14] is based on the growth estimate of f' , which we discuss in the next section, and on an idea of Carleson (see [4]) that if a disc has a large harmonic measure, then the harmonic measure of an arc with endpoints in this disc is also large. Recently, Rohde [32] has found an elegant elementary way to avoid the latter argument.

We consider next two special cases of the theorem—both actually admitting simpler proofs. The first one concerns the Hausdorff dimension of ω . It is now well known (see [6, 11]) that for any $\varepsilon > 0$, $|f'(r\zeta)|^{\pm 1} = o((1-r)^{-\varepsilon})$ as $r \rightarrow 1$ a.e. on $\partial\mathbf{D}$. Consequently, $\dim \omega = 1$ for any simply-connected domain which improves an earlier estimate of Carleson [4]: $\dim \omega \geq \text{const} > \frac{1}{2}$. (Note that the inequality $\dim \omega \geq \frac{1}{2}$ is almost immediate.) It should be noted here that Carleson's proof is much more interesting than his estimate; in §3 we shall apply it to another problem on boundary distortion.

The second case concerns relations with the linear measure Λ_1 . By choosing $\phi(t) = t$ we get the following:

$$\omega \ll \Lambda_1 (\perp \Lambda_1) \Leftrightarrow f' \text{ exists (fails to exist) a.e. on } \partial\mathbf{D}. \quad (3)$$

This may be considered as an extension of (a part of) the F. and M. Riesz theorem: $\omega \ll \Lambda_1$ on a rectifiable boundary. Certainly, there are domains with nonrectifiable boundaries such that f' exists a.e. on $\partial\mathbf{D}$ (e.g., when $\partial\Omega$ has no twist points (see [19]); in particular, when Ω is starlike). On the other hand,

there are univalent functions f satisfying $\liminf_{r \rightarrow 1} |f'(r\zeta)| = 0$ a.e. on $\partial\mathbf{D}$, and, moreover, recently Jones constructed a function with a stronger property: $f'(r\zeta) \rightarrow 0$ in measure on $\partial\mathbf{D}$ as $r \rightarrow 1$ (see [27]). It was Pommerenke who gave the following nice description for this type of singularity. Let $p_\varepsilon(e)$ denote the ε -entropy of a set e , i.e., the minimal number of sets of diameter ε needed to cover e . If $f'(r\zeta) \rightarrow 0$ in measure on $\partial\mathbf{D}$, then there are sets $e_\varepsilon \subset \partial\Omega$ satisfying

$$\omega(e_\varepsilon) \rightarrow 1 \text{ and } \varepsilon p_\varepsilon(e_\varepsilon) \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0, \quad (4)$$

the converse being true for quasidisks. Of course, (4) is stronger than $\omega \perp \Lambda_1$.

In [26] Pommerenke applied a local version of (3) to give a very simple proof of the fact that ω is supported on a set of σ -finite linear measure. It is remarkable that this result extends to general nonsimply-connected domains, but the proof of Jones and Wolff [9] is far more difficult.

Finally, we mention a beautiful relevant result of Bishop, Carleson, Garnett, and Jones (to appear). Note first that (3) implies that $\omega \not\perp \Lambda_1$ iff there is a rectifiable simple curve γ in the closure $\overline{\Omega}$ with $\Lambda_1(\partial\Omega \cap \gamma) > 0$. Now let $\partial\Omega$ be a Jordan curve, and ω^+ and ω^- denote the harmonic measures of the complementary components $\Omega^+ = \Omega$ and Ω^- . Then $\omega^+ \not\perp \omega^-$ iff there are rectifiable curves γ^+ and γ^- lying in $\overline{\Omega}^+$ and $\overline{\Omega}^-$ respectively such that

$$\Lambda_1(\gamma^+ \cap \gamma^-) > 0.$$

2. The growth of the derivative: the LIL estimate. Further information about harmonic measure is derived by a more careful study of the behavior of f' on almost all radii. We consider the corresponding problem for Bloch functions, i.e., for holomorphic functions b on \mathbf{D} satisfying $\|b\|_{\mathcal{B}} = b(0) + \sup(1 - |z|^2)|b'(z)| < \infty$. It is well known that $\|\log f'\|_{\mathcal{B}} \leq 6$ for any univalent function f with $f'(0) = 1$, and conversely $\|\log f'\|_{\mathcal{B}} \leq 1$ and $f'(0) = 1$ imply the univalence of f . Our estimate has the form of the upper class part of the law of the iterated logarithm (LIL). Somewhat weaker versions were obtained by Clunie and MacGregor [6] and Korenblum [11].

THEOREM 2. *There exists an absolute constant C such that if b is a Bloch function, then for almost all $\zeta \in \partial\mathbf{D}$,*

$$\limsup |b(r\zeta)| / \left(\log \frac{1}{1-r} \log \log \log \frac{1}{1-r} \right)^{1/2} \leq C \|b\|_{\mathcal{B}} \quad \text{as } r \rightarrow 1. \quad (5)$$

The original proof in [13] is rather awkward but it explains, to some extent, the probabilistic nature of the result. It is known that $\|b\|_{\mathcal{B}} \asymp \sup_{n \geq 0} \|b_n\|_{\infty}$, where b_n denotes the convolution of b with the polynomial W_n whose coefficients $\hat{W}_n(k)$ are zero outside $(2^{n-1}, 2^{n+1})$ and linear on $[2^{n-1}, 2^n]$ and on $[2^n, 2^{n+1}]$ with $\hat{W}_n(2^n) = 1$. The idea of proving (5) is just to use the weak dependence of the polynomials b_n . A better understanding comes from a subsequent paper of Przytycki [30] in which he observes that (5) is equivalent to the upper class

estimate

$$\limsup_{n \rightarrow \infty} \sum_{k=1}^n b_k / (n \log \log n)^{1/2} \leq \text{const} \|b\|_{\mathcal{B}}$$

and explains how a standard procedure [25] reduces it to the LIL for martingales.

At the same time it was observed (see [25]) that the analytic argument of Clunie and MacGregor [6], based on the Hardy identity, gave a simpler proof of (5) with $C = 1$. This value of C is nearly optimal. Pommerenke [25] observed that the lacunary series $b(z) = \sum z^{15^k}$ satisfies, by the LIL result of M. Weiss [33], the lower class estimate: \limsup in (5) $\geq 0.685 \|b\|_{\mathcal{B}}$ on almost all radii.

All of this has the following consequence to harmonic measure. For $c > 0$ let ϕ_c denote the function $t \mapsto \exp\{c(\log t^{-1} \log \log \log t^{-1})^{1/2}\}$. There exists an absolute constant $c > 0$ such that $\omega \ll \Lambda_{\phi_c}$ for any simply-connected domain, and there is a Jordan domain satisfying $\omega \perp \Lambda_{\phi_c}$ for some $c > 0$. (See [13].)

The second assertion refines earlier results of Carleson [4], and Kaufman and Wu [10]. It is worth mentioning that the research in this direction was originated by Lavrentiev's geometrical example of 1936 of a Jordan domain with $\omega \not\ll \Lambda_1$ [12]. The abovementioned proof of the existence of Ω with maximally singular ω relies on a univalence criterion and hence gives no idea of the geometry of the domain. Subsequently, it was proved (see [15, 31]) that the von Koch snowflake, a domain defined by a geometrical construction, also satisfies $\omega \perp \Lambda_{\phi_c}$ and, moreover, almost every simply-connected domain with a scale invariant structure has this property (see [31]). These examples will be discussed in §5 within the framework of dynamical systems' approach.

A few months ago, Jones proved a very powerful theorem on the validity of $\omega \perp \Lambda_{\phi_c}$. His sufficient condition consists, roughly speaking, of the wiggling of the boundary on every scale at every boundary point. It is fulfilled for all domains like the snowflake.

An interesting result of a different nature was established by Rohde [32]. He showed that the ε -entropy counterpart of the problem considered has no connection with the LIL: if $e \subset \partial\Omega$ and ε is small enough, then

$$p_{\varepsilon}(e) \geq \varepsilon^{-1} \exp \left\{ -c \left(\log \frac{1}{\varepsilon} / \omega(e) \right)^{1/2} \right\}.$$

It is not clear how precise this estimate is.

3. Distortion of Hausdorff dimension. The results considered in this section, although they are not being directly connected with our main topic, shed more light on metric properties of the boundary distortion. Let f be a univalent function mapping \mathbf{D} onto Ω . What can we say about the size of $e \subset \partial\Omega$ if it is known that the set $f^{-1}e$ has dimension p , $0 < p \leq 1$? We shall concentrate on lower bounds for $\dim e$; no nontrivial upper bound is possible—a set of zero dimension may be mapped onto a set of positive area (see [18]). For $p \in (0, 1]$ let $d(p)$ denote the largest number d satisfying $\dim f^{-1}e = p \Rightarrow \dim e \geq d$ for any f . Since $|f'(z)| \geq \text{const}(1 - |z|)$ on \mathbf{D} , we have $p/2 \leq d(p) \leq p$. For $p = 1$

we have $d(1) = 1$, which is an extension of the estimate $\dim \omega \geq 1$ (see §1), but it turns out that for all other values of p both trivial bounds can be improved. The reason is that for any $p < 1$ (but not for $p = 1$) f' may have a positive order of growth on a set of dimension p . On the other hand, Carleson's ingenious argument in the proof of the estimate $\dim \omega \geq \text{const} > \frac{1}{2}$ extends to any $p > 0$ and gives the strict inequality $d(p) < p$.

THEOREM 3. *If $p \in (0, 1]$, then $p/2 < d(p) \leq p(2-p)^{-1}$. Furthermore, $p - d(p) \asymp \sqrt{1-p}$ as $p \rightarrow 1$. Consequently, $\lim_{p \rightarrow 1} p^{-1}d(p) = 1$ and $\lim_{p \rightarrow 0} p^{-1}d(p) = \frac{1}{2}$.*

Upper bounds for $d(p)$ are obtained by constructing the corresponding examples. In proving that $d(p) \leq p(2-p)^{-1}$, one may apply starlike functions; the derivative has an explicit description and is easily studied in this case. For p close to one, we prove the stronger estimate $d(p) \leq p - \text{const}\sqrt{1-p}$ by considering univalent functions f with $\log f'$ being a lacunary series. A relevant property of such functions is that $\int |f'(r\zeta)|^\delta |d\zeta| \geq \text{const}(1-r)^{-c\delta^2}$ as $\delta \rightarrow 0$. The converse estimate

$$\int |f'(r\zeta)|^\delta |d\zeta| = O((1-r)^{-c\delta^2}) \quad (6)$$

holds for any univalent function and plays an important role in various problems on conformal mappings (see [23]). In particular, (6) implies the lower bound $d(p) \geq p - \text{const}\sqrt{1-p}$ and hence the inequality $d(p) > p/2$ for p close to one. Recently, Pommerenke [24] obtained a refined version of (6) which gives $d(p) > p/2$ for $p > 0.601$, but for smaller values Carleson's device remains to be the only known method yielding this result.

4. Nonsimply-connected case. For a general planar domain Ω and a point $a \in \Omega$, the harmonic measure $\omega = \omega(\cdot; \Omega, a)$ evaluated at a is defined by the requirement that for any continuous function u on $\partial\Omega$, $\int u d\omega$ is the value at a of the solution of the Dirichlet problem with boundary values u . Øksendal proved in [20, 21] that ω is always singular with respect to area measure and conjectured that $\omega \perp \Lambda_p$ for all $p > 1$. As we noted in §1, the proof of this conjecture for simply-connected domains, as well as Pommerenke's proof of the refinement that ω is supported on a set of σ -finite Λ_1 -measure, easily follows from the connection with conformal mapping. It is also easy to see that both proofs work with minor changes for domains satisfying the following Milloux type condition:

$$\begin{aligned} &\text{If } a \in \Omega \text{ is sufficiently close to } \partial\Omega \\ &\text{and } D = \{|z - a| < 2\text{dist}(a, \partial\Omega)\}, \\ &\text{then } \omega(\partial\Omega \cap D; \Omega \cap D, a) \geq \text{const} > 0, \end{aligned} \quad (7)$$

but it is far more difficult to dispense with this condition. Øksendal's conjecture had been open for several years until finally Jones and Wolff obtained the following outstanding result.

THEOREM 4 (JONES AND WOLFF). *Let Ω be a planar domain and $a \in \Omega$. Let*

$$e = \{\zeta \in \partial\Omega: \limsup_{r \rightarrow 1} r^{-1} \omega(\{|z - \zeta| < r\}; \Omega, a) > 0\}.$$

Then e has σ -finite linear measure and $\omega(e) = 1$.

The proof of this theorem is very involved. Apart from a hard theoretic-potential technique and a rather sophisticated combinatorial argument, it uses an idea from Carleson's paper [5] to apply the Poisson-Jensen formula

$$\frac{1}{2\pi} \int \left(\log \frac{\partial g}{\partial n} \right) \frac{\partial g}{\partial n} = \gamma + \sum_{\nabla g(z)=0} g(z) \quad (8)$$

involving critical points of Green's function g and the Robin constant γ . If the domain is very nonsimply-connected, then g has many critical points and the sum to the right is large. In [5] Carleson used this fact to show that for square Cantor sets, $\dim \omega$ is less than one. We shall discuss that fundamental paper of Carleson in the next section and here we state another remarkable theorem of Jones and Wolff generalizing Carleson's result to arbitrary "Cantor-like" sets without a scale invariant structure.

THEOREM 5 (JONES AND WOLFF). *Suppose Ω satisfies (7) and (for some $\varepsilon > 0$) the following condition: For every $\zeta \in \partial\Omega$ and $r \leq \text{diam } \partial\Omega$, there is an open topological annulus in $\Omega \cap \{r < |z - \zeta| < \varepsilon^{-1}r\}$ which is conformally equivalent to an annulus containing $1 - \varepsilon < |z| < 1$. Then $\dim \omega < 1$.*

We refer to the paper [8] of Jones and Marshall for some other important results concerning the harmonic measure of nonsimply-connected domains.

5. Scale invariant case. In dynamical systems there has been interest in the question on the Hausdorff dimension of a measure invariant under a transformation. It has also been observed [5, 17, 28, 29] that the corresponding methods could be applied to study harmonic measure in certain scale invariant cases. In the presence of dynamics or selfsimilarity, these methods provide simpler proofs to most of the results discussed above and even enable us to improve some of them. In [5] Carleson outlined a dynamical approach to harmonic measure on Cantor sets and snowflake curves; here we shall discuss a refinement of Carleson's approach (see [15, 16]). Independently, Przytycki [28, 29] applied a somewhat different approach to study harmonic measure on the boundary of a simply-connected attractive basin for a holomorphic map; the subsequent significant paper [31] extended his ideas to very general "fractal" curves. With both approaches, the key idea is to introduce a Markov map F defined on $\partial\Omega$ and to express ω , or a measure equivalent to ω , as a certain Gibbs equilibrium measure invariant under F . We refer to [1] for the relevant notions and terminology of symbolic dynamics.

We illustrate this idea by the example of Cantor type mixing repellers. These are the sets J defined by the following construction. Let Q_1, \dots, Q_m be conformal discs with disjoint closures and F be a function univalent on each Q_i . Let

\sum_m denote the space of infinite sequences $x = (x_1 x_2 \cdots)$ of the symbols $1, \dots, m$ and let $\Delta = (\Delta_{ij})$ be an $m \times m$ matrix consisting of zeros and ones such that $\Delta_{ij} = 1$ implies $\bar{Q}_j \subset FQ_i$. We define $\sum(\Delta) = \{x \in \sum_m: \Delta_{x_i x_{i+1}} = 1\}$ and assume that the shift $T: (x_1 x_2 \cdots) \mapsto (x_2 x_3 \cdots)$ is mixing on $\sum(\Delta)$. We consider nonzero cylinder sets $X = (x_1 \cdots x_n) = \{y \in \sum(\Delta): y_1 = x_1, \dots, y_n = x_n\}$ and call $n = |X|$ the rank of X . For each X of rank $n \geq 2$ we define $Q_X = (F|Q_{x_1})^{-1} \cdots (F|Q_{x_{n-1}})^{-1} Q_{x_n}$ and set $J_n = \bigcup_{|X|=n} \bar{Q}_X$. Finally, we define $J = \bigcap J_n$. Usual Cantor sets, the limit sets of Schottky groups, and some Julia sets provide many examples. It is easy to see that J is invariant under F and that $\text{diam } Q_X \leq \text{const } q^{|X|}$ for some $q < 1$. Thus there is a 1-to-1 correspondence between the points of the sets J and $\sum(\Delta)$ so that $F|J$ corresponds to the shift. We shall identify these dynamical systems; in particular, we shall write X for $J \cap Q_X$ and consider the harmonic measure ω evaluated at ∞ both on J and $\sum(\Delta)$.

The basic property of Cantor type mixing repellers is that the relative harmonic measure is almost scale invariant (see [5, 16]):

$$\left| \frac{\omega(XYZ)}{\omega(XY)} : \frac{\omega(YZ)}{\omega(Y)} - 1 \right| \leq \text{const } q^n \quad (n = |Y|). \quad (9)$$

The reason comes from the fact that the sets XYZ and XY are separated from ∂Q_X by n topological annuli of modules $\geq \text{const}$. From (9) it follows that there is a unique shift invariant probability measure μ on $\sum(\Delta)$ which is equivalent (with bounded densities) to ω . (For μ , one can take any *-weak limit point of the sequence of the measures $n^{-1} \sum_{k=1}^n \omega_k$, where ω_k is given by $\omega_k(X) = \sum_{|A|=k} \omega(AX)$.) Then μ has a property similar to (9) and hence for any $x \in \sum(\Delta)$ there exists a limit $\varphi(x) = \lim \log[\mu(x_1 \cdots x_n)/\mu(x_2 \cdots x_n)]$, and φ satisfies the Hölder condition. Since

$$\mu(x_1 \cdots x_n) \asymp \exp \left\{ \sum_{i=1}^{n-1} \varphi(T^i x) \right\},$$

μ is the Gibbs measure for φ . For $\dim \mu$ we have the following formula in which h_μ denotes the entropy of μ and g the Green's function of $\Omega = \hat{\mathbb{C}} \setminus J$ with pole at infinity.

THEOREM 6.

$$\dim \omega = \dim \mu = \left(1 + h_\mu^{-1} \lim_{n \rightarrow \infty} n^{-1} \sum_{\nabla g(z)=0, z \notin J_n} g(z) \right)^{-1}$$

where the limit always exists and is positive.

The proof (see [16]) closely follows Carleson's reasoning in [5] and is based on the application of (8) and the Manning formula $\dim \mu = h_\mu/\chi_\mu$, where χ_μ denotes the Lyapunov characteristic exponent $\int \log |F'| d\mu$. In particular we have $\dim \omega < 1$, and it was this way that Carleson obtained the first result of this type.

Theorem 6 is especially clear in the case of polynomial dynamics. Let $F(z) = z^m + \dots$ be a polynomial such that the orbits of its critical points z_1, \dots, z_{m-1} (the zeros of F') tend to ∞ . By the Fatou structural theorem (see [3, §10]), the Julia set J of F is a Cantor type mixing repeller. Also it is known (see [3]) that ω is F -invariant, i.e., $\mu = \omega$, that $h_\omega = \log m$, and that $\text{cap } J = 1$. Hence

$$\begin{aligned}\chi_\mu &= \int \log |F'| d\omega = \log m + \sum_{i=1}^{m-1} \int \log |z - z_i| d\omega(z) \\ &= \log m + \sum_{i=1}^{m-1} g(z_i).\end{aligned}$$

It is easy to verify that the limit in Theorem 6 equals $\sum_{i=1}^{m-1} g(z_i)$ and the assertion follows.

Further information may be obtained by studying the asymptotic variance

$$\sigma^2 = \lim_{n \rightarrow \infty} n^{-1} \int \left(\sum_{j=0}^{n-1} \psi \circ T^j \right) d\mu$$

for the Hölder continuous function $\psi = \varphi + \kappa \log |F'|$, where κ denotes $\dim J$. (Remark: $\int \psi d\mu = 0$). By properties of Gibbs measures, σ^2 is finite and hence the following dichotomy occurs (see [31, 16]): either σ^2 is zero, which is equivalent, by the variational principle, to the case $\dim \omega = \dim J$, or $\sigma^2 > 0$. In the former case, ω is equivalent to Λ_κ , whereas in the latter case standard almost sure invariance principles (see [22]) are valid. Namely, Kolmogorov's test along with the usual covering technique (cf. [2]) shows that if $\sigma^2 > 0$, then $\omega \ll \Lambda_\phi$ or $\omega \perp \Lambda_\phi$ with $\phi(t) = t^\kappa \exp\{h(\log \frac{1}{t})\}$, as $\int^\infty t^{-3/2} h(t) \exp\{-\frac{1}{2}\sigma^{-2}\chi_\mu t^{-1} h^2(t)\} dt$ converges or diverges, respectively.

Here is a simple example illustrating the case $\sigma^2 > 0$. Let J be the linear Cantor set of constant ratio a , $0 < a < \frac{1}{2}$. We shall use the natural coding of J with the symbols 1 and 2. Consider the cylinders $X_n = (1 \dots 1)$ and $Y_n = (1 \dots 12)$ of rank n . We must show that ω is not equivalent to a Hausdorff measure, and it is enough to check that $\omega(X_n) \geq (1 + \delta)\omega(Y_n)$ with $\delta > 0$ independent of n , $n \geq 2$. We have $\omega(X_n) - \omega(Y_n) = \int_{J \setminus X_n} (v - u) d\omega$, where u and v denote the harmonic measures of X_n and Y_n with respect to $\hat{C} \setminus X_{n-1}$. Clearly, $v \geq u$ on $J \setminus X_n$ and hence

$$\omega(X_n) - \omega(Y_n) \geq \omega(Y_n) \min_{Y_{n-1}} (v - u) \geq \text{const } \omega(Y_{n-1})$$

because the minimum is scale invariant. Further examples, including arbitrary linear and square Cantor sets, and some Julia sets are given in [16]. On the other hand, I am unaware of any example of a Cantor type mixing repeller satisfying $\sigma^2 = 0$. It seems plausible that this case can never occur.

Besides Cantor sets, Carleson also considered the snowflakes, a class of simply-connected domains with a selfsimilar boundary. As an example we shall examine the von Koch snowflake which is defined by the following construction. To every

side of the equilateral triangle, at the middle, we glue from outside the equilateral triangle three times smaller. To every side of the resulting polygon we glue the equilateral triangle again three times smaller and so on infinitely many times.

The whole machinery developed in the Cantor case will work without changes as soon as we show that there is a coding satisfying (9). Since any approximating polygon has nonequal angles, the natural coding with $\sum(\Delta) = \sum_4$ does not satisfy (9). (In fact, with more efforts, this coding may be used to derive the LIL result; see [17].) It is better to employ the following device due to Przytycki, Urbanski, and Zdunik [31]; see Figure 1.

$$\begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \end{array} \quad \Delta = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Figure 1

Let J_0 denote the one-twelfth part of the boundary and let the piecewise affine transformation $F: J_0 \rightarrow J_0$ be defined so that $\widehat{A_1 A_2}, \widehat{A_3 A_4} \rightarrow J_0$ and $\widehat{A_5 A_6} \rightarrow \widehat{A_3 A_6}$ preserving orientation, and $\widehat{A_2 A_3} \rightarrow J_0$ and $\widehat{A_4 A_5} \rightarrow \widehat{A_3 A_6}$ reversing it. Then (9) is valid because for any cylinder X , the structure of the snowflake is the same in neighborhoods of X and FX (cf. [5, 15]). Thus we may apply the theory described above to obtain the following result of Przytycki, Urbanski, and Zdunik (we have $\sigma^2 > 0$ because the boundary of the snowflake is not rectifiable): There is a number $c_0 > 0$ such that

$$\omega \perp \Lambda_{\phi_c} \text{ or } \ll \Lambda_{\phi_c} \text{ as } c \leq c_0 \text{ or } c > c_0, \text{ respectively,}$$

where $\phi_c(t) = t \exp\{c(\log \frac{1}{t} \log \log \log \frac{1}{t})^{1/2}\}$.

The paper [33] contains a lot of other interesting results concerning the simply-connected case. In conclusion, we cite one of them connected with Julia sets.

THEOREM 7 (PRZYTICKI, URBANSKI, AND ZDUNIK). *Let Ω be a simply-connected domain and F a holomorphic map defined on a neighborhood U of $\partial\Omega$ such that $F(U \cap \Omega) \subset \Omega$, $F(\partial\Omega) = \partial\Omega$, and $\bigcap_{n \geq 0} F^{-n}(\bar{\Omega} \cap U) = \partial\Omega$. Then there is a nonnegative number c_0 such that $\omega \perp \Lambda_{\phi_c}$ for every $c < c_0$, and $\omega \ll \Lambda_{\phi_c}$ for every $c > c_0$. Furthermore, in case $F|_{\partial\Omega}$ is expanding, $\omega \perp \Lambda_{\phi_{c_0}}$ and also we have $c_0 = 0$ iff $\partial\Omega$ is a real-analytic curve.*

By definition, $F|_{\partial\Omega}$ is expanding if $F'_n \geq \text{const } \lambda^n$ on $\partial\Omega$ with $\lambda > 1$, where F'_n denotes the n th iterate of F . This case occurs, for example, when $\partial\Omega$ is the Julia set of the polynomial $z^2 + a$ with a small enough. In the expanding case, it is not hard to see that ω is again equivalent to a Gibbs measure for a Markov

partition on $\partial\Omega$, and so $c_0 = 0$ iff $\partial\Omega$ is rectifiable. The authors show that, in the expanding case, the latter implies that $\partial\Omega$ is a real-analytic Jordan curve; the first result of this type was established by Fatou (see [7]). Of course, the most difficult part of Theorem 7 is the part concerning the nonexpanding case, and the proof is technically much harder. The authors conjecture that, even without the assumption that $F|_{\partial\Omega}$ is expanding, $c_0 = 0$ iff $\partial\Omega$ is an analytically embedded circle or interval. The Julia set of $z^2 - 2$ illustrates the nonexpanding case with $c_0 = 0$.

REFERENCES

1. R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math., Vol. 470, Springer-Verlag, Berlin and New York, 1975.
2. —, *Hausdorff dimension of quasi-circles*, Inst. Hautes Études Sci. Publ. Math. **50** (1979), 11–26.
3. H. Brolin, *Invariant sets under iteration of rational functions*, Ark. Mat. **6** (1965), 103–144.
4. L. Carleson, *On the distortion of sets on a Jordan curve under conformal mapping*, Duke Math. J. **40** (1973), 547–559.
5. —, *On the support of harmonic measure for sets of Cantor type*, Ann. Acad. Sci. Fenn. Ser. A I Math. **10** (1985), 113–123.
6. J. Clunie and T. MacGregor, *Radial growth of the derivative of univalent functions*, Comment. Math. Helv. **59** (1984), 362–375.
7. P. Fatou, *Sur les équations fonctionnelles*, Bull. Soc. Math. France **48** (1920), 208–314.
8. P. W. Jones and D. E. Marshall, *Critical points of Green's function, harmonic measure, and the corona problem*, Ark. Mat. **23** (1985), 281–314.
9. P. W. Jones and T. H. Wolff, *Hausdorff dimension of harmonic measure in the plane*, Preprint, 1986.
10. R. Kaufman and J.-M. Wu, *Distortion of the boundary under conformal mapping*, Michigan Math. J. **29** (1982), 267–280.
11. B. Korenblum, *BMO estimates and radial growth of Bloch functions*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), 99–102.
12. M. A. Lavrentiev, *Boundary problems in the theory of univalent functions*, Mat. Sb. **1** (1936), 815–846; English transl. in Amer. Math. Soc. Transl. **32** (1963), 1–35.
13. N. G. Makarov, *On the distortion of boundary sets under conformal mappings*, Proc. London Math. Soc. **51** (1985), 369–384.
14. —, *Conformal mapping and Hausdorff measures*, Ark. Mat. (to appear).
15. —, *On the harmonic measure of the snowflake*, Preprint, LOMI E-4-86, Leningrad, 1986.
16. N. G. Makarov and A. L. Volberg, *On the harmonic measure of discontinuous fractals*, Preprint, LOMI E-6-86, Leningrad, 1986.
17. A. Manning, *The dimension of the maximal measure for a polynomial map*, Ann. of Math. **119** (1984), 425–430.
18. K. Matsumoto, *On some boundary problems in the theory of conformal mappings of Jordan domains*, Nagoya Math. J. **24** (1964), 129–141.
19. J. E. McMillan, *Boundary behaviour of a conformal mapping*, Acta Math. **123** (1969), 43–67.
20. B. Øksendal, *Null sets for measures orthogonal to $R(X)$* , Amer. J. Math. **94** (1972), 331–342.
21. —, *Brownian motion and sets of harmonic measure zero*, Pacific J. Math. **95** (1981), 179–192.
22. W. Philipp and W. Stout, *Almost sure invariance principles for partial sums of weakly dependent random variables*, Mem. Amer. Math. Soc. No. 161 (1975).

23. Ch. Pommerenke, *Univalent functions*, Vandenhoeck & Ruprecht, Göttingen, 1975.
24. —, *On the integral means of the derivative of a univalent function*, J. London Math. Soc. **32** (1985), 254–258.
25. —, *The growth of the derivative of a univalent function*, The Bieberbach Conjecture: Proc. Symposium on the Occasion of the Proof, Math. Surveys Monogr. No. 21, Amer. Math. Soc., Providence, R. I., 1986, pp. 143–152.
26. —, *On conformal mapping and linear measure*, J. Analyse Math. **46** (1986), 231–238.
27. —, *On graphical representation and conformal mapping*, Preprint, TU, Berlin, 1986.
28. F. Przytycki, *Hausdorff dimension of harmonic measure on the boundary of an attractive basin for a holomorphic map*, Invent. Math. **80** (1985), 161–179.
29. —, *Riemann map and holomorphic dynamics*, Invent. Math. **85** (1986), no. 3.
30. —, *On the law of iterated logarithm for Bloch functions*, Preprint, Univ. of Warwick, 1986.
31. F. Przytycki, M. Urbanski, and A. Zdunik, *Harmonic, Hausdorff and Gibbs measures on repellers for holomorphic maps*, Preprint, 1986.
32. S. Rohde, *On an estimate of Makarov in conformal mapping*, Preprint, TU, Berlin, 1986.
33. M. Weiss, *The law of the iterated logarithm for lacunary trigonometric series*, Trans. Amer. Math. Soc. **91** (1959), 444–469.

LENINGRAD BRANCH, STEKLOV INSTITUTE OF MATHEMATICS, ACADEMY OF SCIENCES OF THE USSR, 191011 LENINGRAD, USSR

The Geometry of Deformations of a Riemann Surface

SCOTT A. WOLPERT

A major theme of the deformation theory of surfaces is the analogy of the Teichmüller space T_g + mapping class group Γ_g with a symmetric space $K \backslash G$ + arithmetic group Γ . An example is found in the result of Harer-Zagier [8] and Penner [11] that the Euler characteristic of $\Gamma_{g,1}$ is given by the Riemann zeta function, $\zeta(1-2g)$ ($T_{g,n}$ is the genus g , n punctured Teichmüller space, and $\Gamma_{g,n}$ the mapping class group), an obvious parallel to the result of Harder for arithmetic groups [6]. A second example is given by the result (see §2) that $M_g = T_g/\Gamma_g$ has a compactification \overline{M}_g , that is a projective variety. The analogy is to the result of Baily for G of Hermitian type, Γ arithmetic that $\overline{K \backslash G/\Gamma}$ is projective. And just recently F. Bonahon has discovered a description of T_g as a subset of the hyperquadric in an infinite-dimensional Lorentz space (see §3). A second theme of the geometry has always been the interplay between the hyperbolic geometry of a surface and the differential geometry of Teichmüller space.

Our plan is to sketch the results on three questions that have occupied our interest. This is not a survey; proper references can be obtained from the papers cited here.

We recall the basic definitions. Let M be a compact surface (without metric) of genus $g > 1$. Consider a pair $(R, [f])$ where $[f]$ is the homotopy class of a homeomorphism

$$f: M \rightarrow R, \quad R \text{ a Riemann surface}$$

(with hyperbolic metric). Pairs $(R, [f])$ and $(S, [h])$ are defined to be equivalent provided $[h \circ f^{-1}]$ contains a conformal representative. T_g is the space of equivalence classes $\{(R, [f])\}$. Conceptually consider $\{(R, [f])\}$ as defining a structure on M by pullback.

Recall that T_g is a complex manifold, Γ_g acts properly discontinuously, and thus the quotient $M_g = T_g/\Gamma_g$ is a complex V -manifold. The cotangent space at $R \in T_g$ is naturally isomorphic to $Q(R)$, the space of holomorphic quadratic

Partially supported by the National Science Foundation.

differentials on R (tensors of type $dz \otimes dz$). A Hermitian product

$$\langle \varphi, \psi \rangle = \int_R \varphi \bar{\psi} \Lambda^{-1},$$

where $\varphi, \psi \in Q(R)$ and Λ is the R -hyperbolic area element, is defined for $Q(R)$. The product on the dual is the Weil-Petersson metric for T_g . The W-P metric is Γ_g invariant, Kähler, and not complete [1, 14].

1. Symplectic geometry of twists and lengths. Given a compact Riemann surface R , by the uniformization theorem there exists a unique hyperbolic metric with the assigned conformal structure. The prescribed curvature equation is a second order nonlinear partial differential equation; one does not expect to be able to piece together solutions from subdomains. By contrast, Fenchel-Nielsen found that if the conformal structure is not prescribed a priori then every hyperbolic metric can be obtained by piecing together metrics from simple subdomains.

The construction is based on two observations—first that the lengths λ_1, λ_2 , and λ_3 of alternating sides of a right hexagon (in the hyperbolic plane) can be arbitrarily prescribed in $(0, \infty)$. Now if the hexagon is doubled across the unlabeled sides, a pair of pants P (genus 0, three boundary components) with geodesic boundary of lengths $l_j = 2\lambda_j$, $j = 1, 2, 3$, is obtained. P is the building block for constructing compact hyperbolic surfaces.

Given pants P with boundaries b_j and Q with boundaries β_j , then provided b_1 and β_1 have the same length, P can be attached to Q along b_1 and β_1 . The resulting surface is a geometric sum $P \vee Q$; i.e. $P \vee Q$ has a hyperbolic metric whose restriction to each piece is the original metric. The proof is based on the second observation: the geometry in a tubular neighborhood of width $w(l) > 0$ of a simple closed geodesic of length l is completely determined by l .

By combining $2g - 2$ pairs of pants, a compact surface of genus g can be constructed. Fenchel-Nielsen found that by varying the parameters of the construction, every compact hyperbolic metric is obtained. Indeed there are two parameters at each attaching site. For P, Q above there is the length $l(b_1) = l(\beta_1)$ and a twist parameter τ (let p_1 on b_1 be the closest point to b_2 and similarly for q_1 on β_1 and β_2 ; define τ to be the distance from p_1 to q_1 along $b_1 \approx \beta_1$). Fenchel-Nielsen found that the parameters (τ_j, l_j) , $1 \leq j \leq 3g - 3$, $\tau_j \in \mathbf{R}$, $l_j \in \mathbf{R}^+$ for a fixed identification pattern \mathcal{P} provide global real-analytic coordinates for Teichmüller space.

We would like to further emphasize two points of the F-N construction.

The geodesic length functions. Given a free homotopy class α for our base surface M , then a basic invariant of a hyperbolic structure R is the length of the unique R -geodesic in the class α .

Twist vector fields. If α is simple, then an associated flow is defined on T_g . Let Δ be an increment of time and $R \in T_g$; denote by α_R the R -geodesic in α . Cut R open along α_R , rotate one boundary relative to the other by an amount Δ , and attach the boundaries to obtain a new hyperbolic metric. The magnitude

Δ parametrizes a flow on T_g ; let t_α be its infinitesimal generator. An example is a F-N coordinate vector field $\partial/\partial\tau_j$.

Let ω be the Weil-Petersson Kähler form and denote Lie derivatives by juxtaposition. We have the following formulas [16, 17]:

$$\text{Duality formula} \quad \omega(t_\alpha, \cdot) = -dl_\alpha; \quad (1)$$

$$\text{Cosine formula} \quad \omega(t_\alpha, t_\beta) = \sum_{p \in \alpha \# \beta} \cos \theta_p; \quad (2)$$

Sine-length formula

$$\begin{aligned} t_\alpha t_\beta l_\gamma = & \sum_{(p,q) \in \alpha \# \gamma \times \beta \# \gamma} \frac{e^{l_1} + e^{l_2}}{2(e^{l(\gamma)} - 1)} \sin \theta_p \sin \theta_q; \\ & - \sum_{(r,s) \in \alpha \# \beta \times \beta \# \gamma} \frac{e^{m_1} + e^{m_2}}{2(e^{l(\beta)} - 1)} \sin \theta_r \sin \theta_s; \end{aligned} \quad (3)$$

$$\text{Canonical coordinates} \quad \omega = \sum_j d\tau_j \wedge dl_j. \quad (4)$$

The right-hand sides of the second and third formulas are evaluated by the trigonometry of the geodesics α, β , and γ on R . For the second formula the sum is over the intersection points p of the geodesics and θ_p is the angle as measured from α to β . For the third formula l_1 and l_2 are the lengths of the segments of γ defined by p and q and similarly for r, s on β defining m_1 and m_2 . Introducing the rescaled twist fields $T_\alpha = (4 \sinh l_\alpha/2)t_\alpha$ we have a simple formula for the bracket

$$[T_\alpha, T_\beta] = \sum_{p \in \alpha \# \beta} T_{\alpha_p \beta^+} - T_{\alpha_p \beta^-} \quad (5)$$

where $\alpha_p \beta^+, \alpha_p \beta^-$ are the two possible curves obtained by replacing the intersection at p with small arcs connecting α and β [16].

Formula (5) is an obvious analogue of the Chevalley result for the structure constants of a semisimple Lie algebra. Formulas (2) and (3) reveal a simple correspondence between the symplectic geometry of the W-P Kähler form and hyperbolic geometry. W. Goldman has found analogues of the above formulas for more general representations of surface groups [5].

We would like to recall two applications of the formulas.

I. The Teichmüller space $T_{1,1}$. The uniformization group of a once punctured torus has a presentation $\{A, B | [A, B] \text{ is parabolic}\}$. The Teichmüller space is a simplex $T_{1,1} = \{(a, b, c) | a + b + c = 1, a, b, c > 0\}$, where $a = \text{tr } A / \text{tr } B \text{ tr } AB$, $b = \text{tr } B / \text{tr } A \text{ tr } AB$, and $c = \text{tr } AB / \text{tr } A \text{ tr } B$ ($\text{tr} = \text{trace}$). The mapping class group has fundamental domain $\{(a, b, c) | a, b, c < 1/2\}$ and $\omega = (da \wedge db)/abc$. By a short calculation $\int_{\mathcal{M}_{1,1}} \omega = \pi^2/6$ [15].

II. Extension of ω to $\bar{\mathcal{M}}_g$, the moduli space of stable curves. Actually the F-N construction applies to a more general class of surface, a Riemann surface with nodes. The original hexagon lengths $\lambda_1, \lambda_2, \lambda_3$ are also allowed to be 0 (for this case the side is replaced by a point on the circle at infinity). P and Q can

be attached along b_1, β_1 with $l(b_1) = l(\beta_1) = 0$ (the points at ∞ are formally identified). The F-N theorem is valid in this context also: every Riemann surface with nodes is obtained by the construction. To introduce local coordinates for $\overline{\mathcal{M}}_g$ a change of variables is needed: replace τ_j by the angle $\theta_j = 2\pi\tau_j/l_j$ (values mod 2π). The parameters (θ_j, l_j) are polar coordinates (θ_j is undefined when $l_j = 0$) and by formula (5) $\omega = \sum_j l_j d\theta_j \wedge dl_j$. By allowing the l_j to vanish we have local coordinates for $\overline{\mathcal{M}}_g$ (the moduli space of noded Riemann surfaces = stable curves). Extending ω to $\overline{\mathcal{M}}_g$ is equivalent, by the above formula, to defining $r dr d\theta$ at the origin.

2. Obtaining an embedding from the W-P class. Bailly has shown that for G of Hermitian type, $V = K \backslash G/\Gamma$ has a compactification \overline{V} with $V \subset \overline{V}$ open, dense and \overline{V} has a projective embedding [2]. The analogue for \mathcal{M}_g is true. $\overline{\mathcal{M}}_g$ is a complex V -manifold, $\mathcal{D} = \overline{\mathcal{M}}_g - \mathcal{M}_g$ is a divisor, and $\overline{\mathcal{M}}_g$ has a projective embedding [10]. We shall describe a proof by analytic methods [18].

The proof is suggested by the applications I and II. The extension of ω defines a symplectic form on $\overline{\mathcal{M}}_g$. Provided ω determines a rational class in $H^2(\overline{\mathcal{M}}_g)$, the existence of an embedding follows from the Kodaira theorem.

Testing for rationality requires integrating ω over a large number of 2-cycles ($H_2(\overline{\mathcal{M}}_g, \mathbb{Q})$ has rank $2 + [g/2]$). At the outset this appears to be a formidable task. Two formal properties allow one to reduce to the evaluation of the single integral $\int_{\mathcal{M}_{1,1}} \omega$.

In lieu of the argument, we shall describe one of the formal properties: the restriction phenomenon. Suppose $\mathcal{P} = \{\gamma_j\}_{j=1}^{3g-3}$ is the collection of $3g - 3$ disjoint geodesics where the pants were joined to obtain a surface R with nodes. Assume γ_k separates R into two components: R_1 containing $\{\gamma, \dots, \gamma_{k-1}\}$ and R_2 containing $\{\gamma_{k+1}, \dots, \gamma_{3g-3}\}$. Let \mathcal{L} be the subset of $\overline{\mathcal{M}}_g$ described in F-N coordinates by $l_k = 0$. Our discussion suggests the description of \mathcal{L} as the product of the moduli spaces of R_1 and R_2 (the τ_j, l_j are independent of one another). Formula (5) shows $\omega(R) = \omega(R_1) + \omega(R_2)$; the Kähler form on \mathcal{L} is the sum of two lower genera forms. With this property we are able to reduce the genus of the calculation to the one case: genus 1, 1 puncture.

By pursuing this basic idea we are actually able to determine the line bundle that the class $\omega/2\pi^2$ represents [22]. Unfortunately the actual proof is not so simple. Fenchel-Nielsen coordinates and the holomorphic coordinates for $\overline{\mathcal{M}}_g$ are not smoothly equivalent in the directions transverse to \mathcal{D} . A holomorphic parameter t for opening a node is related to the geodesic length by $l \approx 2\pi^2/(\log 1/|t|)$. For the t variable the W-P Kähler form is singular

$$\omega \sim \frac{dt \wedge d\bar{t}}{|t|^2 (\log 1/|t|)^3} = \partial\bar{\partial} \left(\frac{1}{(\log 1/|t|^2)} \right).$$

Nevertheless, ω is the curvature 2-form (in the weak sense) of a continuous metric in a line bundle. This is sufficient for the Kodaira theorem [18].

3. Bonahon's hyperquadric model for T_g . A model for hyperbolic n -space is the hyperquadric $\langle x, x \rangle = -1$, $x_{n+1} > 0$ of Lorentz space ($\langle \cdot, \cdot \rangle$ has signature $(n, 1)$). Francis Bonahon has discovered an analogous model for Teichmüller space and the W-P metric; we sketch the idea below [4].

A homeomorphism $f: R \rightarrow S$, of Riemann surfaces, may be lifted to their universal covers, the hyperbolic plane. The lift extends to S_∞^1 , the circle at infinity. Denote by $\hat{f}: S_\infty^1(R) \rightarrow S_\infty^1(S)$ the restriction; \hat{f} is uniquely determined (a canonical map) given $R, S \in T_g$. Furthermore, since the unit tangent bundle $UT(R)$ of R may be described by triples of points on $S_\infty^1(R)$ the map \hat{f} induces a canonical identification of $UT(R)$ and $UT(S)$, which conjugates (only C°) the geodesic flows. We shall identify the unit tangent bundles $UT(R)$ and geodesic flows $GF(R)$, $R \in T_g$, via the canonical maps.

Consider \mathcal{TM} the linear cone of positive measures on UT transverse to GF . A measure μ is defined for Borel subsets of a 2-plane transverse to GF and is invariant: $\mu(B) = \mu(GF_t(B))$ for all time. An example is given by a free homotopy class α (on M). Let α_R be the geodesic representative on R and $\tilde{\alpha}_R$ its lift to $UT(R)$. Define $\delta_\alpha(B)$ to be the cardinality of $\tilde{\alpha}_R \cap B$; δ_α is an invariant measure. A second example comes from the GF -geometry of the unit tangent bundle. Fix a surface S and let μ_s be its Liouville measure: μ_s is the interior product of the volume form with the infinitesimal generator of $GF(S)$. Alternately $\mu_s(B)$ is the number of intersections per unit length of B and the generic $GF(S)$ trajectory segment. By a theorem of Mostow the map $S \rightarrow \mu_s$ is actually an embedding of T_g into \mathcal{TM} . We shall write \mathcal{T}_g for the image.

A pairing is defined for the measures δ_* : $\delta_\alpha \# \delta_\beta$ is the negative of the minimal number of intersections for representative curves. The span of the measures δ_* is dense in \mathcal{TM} ; the pairing extends to the entire space of transverse measures. Bonahon's analogy is revealed by considering examples for the pairing.

Provided α is simple, then $\delta_\alpha \# \delta_\alpha = 0$; in fact the light cone is simply \mathcal{MGL} , Thurston's space of measured geodesic laminations. Bonahon finds that the image \mathcal{T}_g is asymptotic to the light cone and thus can be compactified by $\{\text{light cone}\}/\mathbf{R}^+$, the sphere at infinity for hyperbolic space. This is precisely Thurston's compactification.

The self-pairing $\mu_R \# \mu_R$ is the negative of the volume of $UT(R)$; \mathcal{T}_g is contained in the hyperquadric $\gamma \# \gamma = 4\pi^2(1 - g)$. The formula $\delta_\alpha \# \mu_R = l_\alpha(R)$ is an exercise for the reader. Bonahon finds an interpretation for the tangent plane at each point of \mathcal{T}_g . Thus it is appropriate to consider the infinitesimal geometry of the model. In particular, if ν is a vector tangent to \mathcal{T}_g at μ_R then indeed $\mu_R \# \nu = 0$ (the position vector is perpendicular to the tangent plane). At this point Bonahon finds a reformulation of a result of the author [20].

THEOREM. *The restriction of $\#$ to the image $\mathcal{T}_g \subset \mathcal{TM}$ is the Weil-Petersson metric.*

To pursue the comparison between the W-P geometry and that of hyperbolic space recall that the metric has negative curvature [12, 13, 19] and that its

exponential maps are homeomorphisms from their domains to T_g [21] (the metric is not complete).

REFERENCES

1. L. V. Ahlfors, *Some remarks on Teichmüller's space of Riemann surfaces*, Ann. of Math. **74** (1961), 171–191.
2. W. L. Baily, *Compactification of arithmetic quotients of bounded symmetric domains*, Ann. of Math. **84** (1966), 442–528.
3. L. Bers, *Spaces of degenerating Riemann surfaces*, Ann. of Math. Stud. **79** (1974), 43–55.
4. F. Bonahon, *Geometry of Teichmüller space via geodesic currents* (in preparation).
5. W. Goldman, *The symplectic nature of fundamental groups of surfaces*, Adv. in Math. **54** (1984), 200–225.
6. G. Harder, *A Gauss-Bonnet formula for a discrete arithmetically defined group*, Ann. Sci. École Norm. Sup. (4) **4** (1971), 409–455.
7. J. Harer, *The cohomology of the moduli space of curves*, preprint.
8. J. Harer and D. Zagier, *The Euler characteristic of the moduli space of curves*, preprint.
9. S. Kerckhoff, *The Nielsen realization problem*, Ann. of Math. **117** (1983), 235–265.
10. F. Knudsen, *The projectivity of the moduli space of stable curves*, Math. Scand. **52** (1983), 161–212.
11. R. Penner, *Perturbative series and the moduli space*, preprint.
12. H. Royden, oral communication.
13. A. J. Tromba, *The curvature of Teichmüller space with respect to its Weil-Petersson metric*, preprint.
14. S. A. Wolpert, *Non-completeness of the Weil-Petersson metric for Teichmüller space*, Pacific J. Math. **61** (1975), 573–577.
15. —, *On the Kähler form of the moduli space of once punctured tori*, Comment. Math. Helv. **58** (1983), 246–256.
16. —, *On the symplectic geometry of deformations of a hyperbolic surface*, Ann. of Math. **117** (1983), 207–234.
17. —, *On the Weil-Petersson geometry of the moduli space of curves*, Amer. J. Math. **107** (1985), 969–997.
18. —, *On obtaining a positive line bundle from the Weil-Petersson class*, Amer. J. Math. **107** (1985), 1485–1507.
19. —, *Chern forms and the Riemann tensor for the moduli spaces of curves*, Invent. Math. (to appear).
20. —, *Thurston's Riemannian metric for Teichmüller space*, J. Differential Geom. **23** (1986), 143–174.
21. —, *Geodesic length functions and the Nielsen problem*, J. Differential Geom. (to appear).
22. —, *On the homology of the moduli space of stable curves*, Ann. of Math. **118** (1983), 491–523.

UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND 20742, USA

Conjugation of Shimura Varieties

M. V. BOROVoi

1. In his address [19] to the 1978 International Congress of Mathematicians in Helsinki, Shimura asked five questions of which the first was: Is it possible to define arithmetic automorphic functions on arithmetic quotients of bounded symmetric domains? The answer “yes” is known as Shimura’s conjecture. The conjecture states that Shimura varieties admit canonical models over small number fields. A Shimura variety in the simplest case is an algebraic variety over \mathbf{C} such that the corresponding analytic space is of the form $\Gamma \backslash X$ where X is a bounded symmetric domain and Γ is an arithmetic group acting on X . Shimura’s conjecture was proved in some cases by Shimura [18], K. Miyake, and K.-y. Shih. The ideas surrounding the notion of canonical models were substantially clarified by Deligne [4, 5], who succeeded in giving a definition of a canonical model in the general case and proving the existence of such models in the “abelian” cases.

In [10] Langlands proposed a remarkable conjecture on the conjugation of Shimura varieties by automorphisms of \mathbf{C} . Langlands’s conjecture implies [14] Shimura’s one. Here we shall state Langlands’s conjecture and prove it for all Shimura varieties, using the already proved Langlands’s conjecture for Shimura varieties of type A_1 .

2. To state Langlands’s conjecture we need the notion of a (connected) pro-algebraic Shimura variety. Shimura varieties are associated with Mumford manifolds.

A *Mumford manifold* is a pair (G, X) , where G is a connected simply connected semisimple \mathbf{Q} -group of Hermitian type and X is the symmetric space of $G(\mathbf{R})$ with an invariant complex structure (“ G is of Hermitian type” means that such a complex structure exists). A Mumford manifold (G, X) is *simple* if G is \mathbf{Q} -simple; any Mumford manifold decomposes into a product of simple ones. A simple Mumford manifold (G, X) is said to be of type A_n, B_n, C_n, D_n, E_6 , or E_7 if G is of this type.

A point $x \in X$ is a *CM-point* if there is a maximal torus $T \subset G$ such that $T(\mathbf{Q}) \cdot x = x$. Such a pair (T, x) is called a *point with complex multiplication* in (G, X) . An *embedding of Mumford manifolds* $(G_1, X_1) \hookrightarrow (G_2, X_2)$ is a pair

of embeddings $f: G_1 \hookrightarrow G_2$, $u: X_1 \hookrightarrow X_2$ such that u is holomorphic, $G_1(\mathbf{R})$ -equivariant, and takes CM-points to CM-points.

A Mumford manifold (G, X) gives rise to a (connected pro-algebraic) Shimura variety $\mathrm{Sh}(G, X) = \lim \mathrm{proj} \Gamma \backslash X$, where Γ runs over the congruence subgroups of $G(\mathbf{Q})$. The set of \mathbf{C} -points of $\mathrm{Sh}(G, X)$ is $(G(\mathbf{A}^f) \times X)/G(\mathbf{Q})$. We denote by $[g, x] \in \mathrm{Sh}(G, X)$ the class of (g, x) , where $g \in G(\mathbf{A}^f)$, $x \in X$. The adèle group $G(\mathbf{A}^f)$ acts on $\mathrm{Sh}(G, X)$ by $g' \cdot [g, x] = [g'g, x]$. An embedding $(G_1, X_1) \hookrightarrow (G_2, X_2)$ gives rise to an embedding $\mathrm{Sh}(G_1, X_1) \hookrightarrow \mathrm{Sh}(G_2, X_2)$ of the corresponding Shimura varieties.

3. Hereafter we fix an arbitrary automorphism τ of \mathbf{C} . Langlands's conjecture completely describes the variety ${}^\tau \mathrm{Sh}(G, X)$ obtained from $\mathrm{Sh}(G, X)$ under the action of τ . Our formulation of Langlands's conjecture differs from the (very beautiful) formulation of Langlands himself (see [10] and also [13, 14]).

Langlands's conjecture, in my interpretation, consists of three statements K , C_1 , and C_2 .

CONJECTURE $K(G, X)$. *There exists a Mumford manifold $({}^\tau G, {}^\tau X)$, an isomorphism $\varphi = \prod \varphi_l: G(\mathbf{A}^f) \rightarrow {}^\tau G(\mathbf{A}^f)$, where $\varphi_l: G(\mathbf{Q}_l) \rightarrow {}^\tau G(\mathbf{Q}_l)$ is algebraic for any prime l , and a φ -equivariant isomorphism $\psi: \tau \mathrm{Sh}(G, X) \rightarrow \mathrm{Sh}({}^\tau G, {}^\tau X)$.*

Conjecture $K(G, X)$ generalizes Kazhdan's theorem [9], which states that $\tau(\Gamma \backslash X) \simeq {}^\tau \Gamma \backslash {}^\tau X$ for some ${}^\tau X$, ${}^\tau \Gamma$.

Let (T, x) be a point with complex multiplication in (G, X) . Assume that $\psi(\tau[e, x]) = [g, x']$, where e is the neutral element, $g \in {}^\tau G(\mathbf{A}^f)$, $x' \in {}^\tau X$. We say that the pair (φ, ψ) is x -normalized if $g = e$. Replacing (φ, ψ) by

$$(\mathrm{int}(g^{-1} \circ \varphi, g^{-1} \circ \psi))$$

we can assume that (φ, ψ) is x -normalized.

Conjecture C_1 describes (G, φ) , provided (φ, ψ) is x -normalized. To see how (G, φ) can be described, we consider the subgroup $\varphi(T(\mathbf{Q})) \subset {}^\tau G(\mathbf{A}^f)$. We have $T(\mathbf{Q}) \cdot x = x$, hence $T(\mathbf{Q}) \cdot [e, x] = [e, x]$ and $\varphi(T(\mathbf{Q})) \cdot [e, x'] = [e, x']$ (recall that $\psi(\tau[e, x]) = [e, x']$). It follows that $\varphi(T(\mathbf{Q})) \cdot [e, {}^\tau X] = [e, {}^\tau X]$, hence $\varphi(T(\mathbf{Q})) \subset {}^\tau G(\mathbf{Q})$. Let $j: T \rightarrow G$ denote the inclusion homomorphism. We see that there exists a homomorphism ${}^\tau j: T \rightarrow {}^\tau G$ such that ${}^\tau j \otimes \mathbf{A}^f = \varphi \circ (j \otimes \mathbf{A}^f)$. In other words, we have constructed a form $({}^\tau G, {}^\tau j: T \rightarrow {}^\tau G)$ of the pair (G, j) and an \mathbf{A}^f -isomorphism of the forms.

Set $T^{\mathrm{ad}} = T/Z$, where Z is the center of G . Forms of (G, j) can be described as isomorphism classes of right tensors (principal homogeneous spaces) of $\mathrm{Aut}(G, j) = T^{\mathrm{ad}}$. With $({}^\tau G, {}^\tau j)$ and φ we associate the torsor $P_\tau = \{\chi: G \xrightarrow{\sim} {}^\tau G \mid {}^\tau j = \chi \circ j\}$ and the \mathbf{A}^f -rational point φ of P_τ . The group T^{ad} acts on P_τ by $\chi \cdot t = \chi \circ \mathrm{int}(t)$. If L/\mathbf{Q} is a finite extension splitting T^{ad} , then P_τ has an L -point.

LEMMA. *Let H be an algebraic torus over a number field k . Let N/k be a Galois extension splitting H . Denote by $\mathcal{P}_k(H)$ the set of isomorphism classes*

of pairs (P, x_f) , where P is a torsor of H and $x_f \in P(\mathbf{A}_k^f)$ (we write $\mathcal{P}(H)$ for $\mathcal{P}_{\mathbf{Q}}(H)$ if $k = \mathbf{Q}$). Then there is a canonical bijection

$$\mathcal{P}_k(H) \simeq (H(\mathbf{A}_N^f)/H(N))^{\text{Gal}(N/k)}.$$

IDEA OF THE PROOF. With (P, x_f) we associate $x_f^{-1} \cdot x_N$, where $x_N \in P(N)$.

Returning to our (T, x) , let L/\mathbf{Q} be a Galois extension splitting T^{ad} . We associate with (P_τ, φ) an element $t \in \mathcal{P}(T^{\text{ad}}) = (T^{\text{ad}}(\mathbf{A}_L^f)/T^{\text{ad}}(L))^{\text{Gal}(L/\mathbf{Q})}$. This element is characterized by the following property: there exists an isomorphism $\chi: G_L \rightarrow {}^\tau G_L$ such that ${}^\tau j = \chi \circ j$ and $\chi = \varphi \circ \text{int}(t)$. Now we can state C_1 .

CONJECTURE $C_1(G, X, x)$. With the above notation, $t = b(T^{\text{ad}}, x, \tau)$ where $b(T^{\text{ad}}, x, \tau)$ is defined in §4.

Let (T_1, x_1) and (T_2, x_2) be two points with complex multiplication in (G, X) . Let $\varphi_1: G(\mathbf{A}^f) \rightarrow {}^\tau G(\mathbf{A}^f)$, $\psi_1: \tau \text{Sh}(G, X) \rightarrow \text{Sh}({}^\tau G, {}^\tau X)$ be an x -normalized pair, i.e., $\psi_1(\tau[e, x_1]) = [e, x'_1]$. We have $\psi_1(\tau[e, x_2]) = [g, x'_2]$, where $g \in {}^\tau G(\mathbf{A}^f)$. Set $\psi_2 = g^{-1} \circ \psi_1$. $\varphi_2 = \text{int}(g^{-1}) \circ \varphi_1$, then (φ_2, ψ_2) is x_2 -normalized. We see that g is a “difference” between ψ_1 and ψ_2 . Conjecture C_2 describes g (see details in §4).

Langlands’s conjecture also describes ${}^\tau X$ and x' (see [3]). We will not dwell on it.

4. Here we give the formulas necessary to state Conjectures C_1 and C_2 . These formulas are not needed for the proof of the conjectures. The proof requires only the functional properties of C_1 and C_2 , listed in the proposition in §6.

We say that (L, F) is a *CM-field* if F is a totally real number field and L is a totally imaginary quadratic extension of F . Suppose $(L, F) \subset (\mathbf{C}, \mathbf{R})$. Set

$$\begin{aligned} \mathbf{S} &= \ker(\text{Nm}_{\mathbf{C}/\mathbf{R}}: R_{\mathbf{C}/\mathbf{R}} G_{m\mathbf{C}} \rightarrow G_{m\mathbf{R}}), \\ {}^L S' &= \ker(\text{Nm}_{L/F}: R_{L/F} G_{mL} \rightarrow G_{mF}), \\ {}^L S &= R_{F/\mathbf{Q}} {}^L S', \end{aligned}$$

where R denotes the restriction of scalars and Nm denotes the norm homomorphism. We have

$$\mathbf{S} = {}^L S' \otimes_F \mathbf{R}.$$

Define an element ${}^L b(\tau) \in \mathcal{P}({}^L S) = \mathcal{P}_F(S') = \{x \in (\mathbf{A}_L^f)^\times / L^\times \mid x \cdot \iota x = 1\}$ where ι is the complex conjugation. Set $\tilde{b} = \tau^{-1} \iota \tau \iota^{-1} \in \text{Aut}(\mathbf{C})$. We have $b \cdot \iota b \iota^{-1} = 1$. Since L is a CM-field, $\tilde{b} \in \text{Aut}(\mathbf{C}/L)$. With the help of the homomorphism $\text{Aut}(\mathbf{C}/L) \rightarrow \text{Gal}(L^{\text{ab}}/L) \xrightarrow{r} \pi_0((\mathbf{A}_L^f)^\times / L^\times)$, where r is the reciprocity law isomorphism, we obtain an element $b_0 \in \pi_0((\mathbf{A}_L^f)^\times / L^\times)$ such that $b_0 \cdot \iota b_0 = 1$. There is a unique element $b \in (\mathbf{A}_L^f)^\times / L^\times$ such that $b \cdot \iota b = 1$ and b_0 is the image of b . Set ${}^L b(\tau) = b$.

For $x \in X$ let K_x^{ad} denote the stabilizer of x in $G^{\text{ad}}(\mathbf{R})$. Consider the homomorphism $h_x: \mathbf{S} \rightarrow K_x^{\text{ad}} \rightarrow G_{\mathbf{R}}^{\text{ad}}$ such that $h_x(z)$ acts in the tangent space $T_x(X)$ as the multiplication by z^2 where $z \in \mathbf{S}(\mathbf{R}) \subset \mathbf{C}$. If (T, x) is a point with complex multiplication, then $\text{im } h_x \subset T_{\mathbf{R}}^{\text{ad}}$. Since $T_{\mathbf{R}}^{\text{ad}}$ is compact, T^{ad} splits

over some CM-field $(L, F) \subset (\mathbf{C}, \mathbf{R})$. Over F , T^{ad} decomposes into a product of one-dimensional tori, hence $h_x: \mathbf{S} = {}^L S' \otimes_F \mathbf{R} \rightarrow T_{\mathbf{R}}^{\text{ad}}$ can be defined over F . We define

$${}^L r: {}^L S = R_{F/\mathbf{Q}} {}^L S' \xrightarrow{R_{F/\mathbf{Q}} h_x} R_{F/\mathbf{Q}} T^{\text{ad}} \xrightarrow{\text{Nm}_{F/\mathbf{Q}}} T^{\text{ad}}.$$

Set $b(T^{\text{ad}}, x, \tau) = {}^L r({}^L b(\tau))$.

REMARK. We can define a pro-algebraic group $S = \lim \text{proj } {}^L S$, an element $b(\tau) = \lim \text{proj } {}^L b(\tau) \in \mathcal{P}(S)$, and a homomorphism $r(x): S \rightarrow T^{\text{ad}}$ such that $t = r(x)(b(\tau))$.

Now let (T_1, x_2) and (T_2, x) be two points with complex multiplication in (G, X) . Let φ_1 , g , and φ_2 be as at the end of §3. Conjecture C_2 describes the image of g in ${}^\tau G(\mathbf{A}^f)/{}^\tau G(\mathbf{Q})$. Let L be a Galois CM-field splitting T_1 and T_2 . It is easy to deduce (see [3]) from C_1 and the formula $\varphi_2 = \text{int}(g^{-1}) \circ \varphi_1$, that the image of g in $G^{\text{ad}}(\mathbf{A}^f)/G^{\text{ad}}(L)$ equals $\varphi_1^{\text{ad}}(r_2({}^L b(\tau)^{-1}) \cdot r_1({}^L b(\tau))) \bmod G^{\text{ad}}(L)$, where $r_1 = {}^L r(x_1)$, $r_2 = {}^L r(x_2)$, and $\varphi_1^{\text{ad}}: G^{\text{ad}}(\mathbf{A}^f) \rightarrow {}^\tau G^{\text{ad}}(\mathbf{A}^f)$ is the isomorphism corresponding to φ_1 . If r_1 and r_2 can be lifted to homomorphisms $\tilde{r}_1, \tilde{r}_2: {}^L S \rightarrow G$, then it is natural to conjecture that

$$g = \varphi_1(\tilde{r}_2({}^L b(\tau))^{-1} \tilde{r}_1({}^L b(\tau))).$$

It turns out that the morphism of varieties $r_2^{-1} r_1: {}^L S \rightarrow G^{\text{ad}}$ can always be uniquely lifted to a certain morphism $\tilde{r}_2^{-1} \tilde{r}_1: {}^L S \rightarrow G$, such that $(\tilde{r}_2^{-1} \tilde{r}_1)(e) = e$.

CONJECTURE $C_2(G, X, x_1, x_2)$. *With the above notation*

$$g = \varphi_1((\tilde{r}_2^{-1} \tilde{r}_1)({}^L b(\tau))).$$

5. Conjecture C_1 implies a simple description of ${}^\tau G$. Suppose (G, X) is simple, $G = R_{F/\mathbf{Q}} G'$, G' being an absolutely simple group over a totally real field F . Then ${}^\tau G = R_{F/\mathbf{Q}} {}^\tau G'$, where ${}^\tau G'$ is the inner form of G' such that ${}^\tau G' \otimes_{F_v} \simeq G' \otimes_{F_v}$ for any nonarchimedean place v of F and that ${}^\tau G' \otimes_{F, \sigma} \mathbf{R} \simeq G' \otimes_{F, \tau^{-1}\sigma} \mathbf{R}$ for every embedding $\sigma: F \hookrightarrow \mathbf{R}$.

In the following section we list some functorial properties of Conjectures C_1 and C_2 .

6. PROPOSITION. *Let (G, X) be a Mumford manifold and (G_0, X_0) be a Mumford submanifold of (G_0, X_0) . Let $x_1, x_2, x_3 \in X_0$ be CM-points. Suppose $\gamma \in G(\mathbf{Q})$. Then*

1. $K(G, X)$ and $C_1(G, X, x_1) \Rightarrow K(G_0, X_0)$ and $C_1(G_0, X_0, x_1)$.
2. *Provided K and C_1 hold,*
 - (a) $C_2(G_0, X_0, x_1, x_2) \Rightarrow C_2(G, X, x_1, x_2)$,
 - (b) $C_2(G, X, x_1, x_2) \Leftrightarrow C_2(G, X, x_1, \gamma x_2)$,
 - (c) $C_2(G, X, x_1, x_2)$ and $C_2(G, X, x_2, x_3) \Rightarrow C_2(G, X, x_1, x_3)$.

The most important statement here is that $K(G, X)$ and $C_1(G, X, x_1)$ implies $K(G_0, X_0)$. This result is actually proved by Milne and Shih [14]. A similar result concerning Shimura's conjecture was earlier proved by Shimura and Deligne. The proof uses the density of the orbit $G_0(\mathbf{A}^f) \cdot [e, x]$ of $[e, x]$ in $\text{Sh}(G_0, X_0)$.

For some Mumford manifolds (for those of abelian type, see below) Shimura's and Langlands's conjectures were proved using the method of symplectic embeddings. This method is due to Shimura and is improved by Deligne [4, 5]. A *symplectic embedding* is an embedding $(G, X) \hookrightarrow (\mathrm{Sp}_n, \mathcal{H}_n)$, where Sp_n is the symplectic group of a $2n$ -dimensional space and \mathcal{H}_n the Siegel space of degree n . A pair (G, X) is said to be of *abelian type* if it admits a symplectic embedding. In this case (G, X) is a moduli manifold of abelian varieties with auxiliary structures (cf. [15]).

For a Mumford manifold (G, X) to be of abelian type it is necessary and sufficient that all its simple factors are of abelian type. A simple Mumford manifold (G, X) is of abelian type if and only if G is of type A_n, B_n, C_n , or of type D_n and $G_{\mathbf{R}}$ has no factors locally isomorphic to the "quaternionic" form $\mathrm{SO}^*(2n)$ of $\mathrm{SO}(2n)$ (see [17, 5]).

If $(G_1, X_1) \hookrightarrow (G_2, X_2)$ is an embedding, then Langlands's conjecture for (G_2, X_2) implies Langlands's conjecture for (G_1, X_1) (see [14]). So to prove Langlands's conjecture in the abelian case, it suffices to prove it for $(\mathrm{Sp}_n, \mathcal{H}_n)$. Milne and Shih [14] showed that Langlands's conjecture for $(\mathrm{Sp}_n, \mathcal{H}_n)$ is equivalent to a remarkable statement about the action of automorphisms of \mathbf{C} on abelian varieties of CM-type and their points of finite order. This statement is proved by Deligne [6]. Thus Langlands's conjecture is proved for all the Mumford manifolds of abelian type—in particular, for those of type A_1 . Earlier, using this method, Deligne [4, 5] proved Shimura's conjecture for all Mumford manifolds of abelian type.

7. If (G, X) is of nonabelian type, then the method of symplectic embeddings cannot be applied. We describe a new method. The first thing that comes into mind is to consider Mumford *submanifolds* of abelian type of (G, X) . However, G can have few subgroups and it is rather difficult to investigate them. Here it is very helpful to use a construction of Piatetski-Shapiro, which enables us to embed any Mumford manifold into a larger one which has many Mumford submanifolds of type A_1 . Using this construction we shall deduce Langlands's conjecture for all Mumford manifolds from Langlands's conjecture for the case A_1 and Conjecture K .

Piatetski-Shapiro's construction. Let (G_0, X_0) be a Mumford manifold and F a totally real field. Set $G' = G_{0F}$, $G = R_{F/\mathbf{Q}}G'$. Then G is of Hermitian type and there is a canonical embedding $G_0 \hookrightarrow G$. This way we construct a Mumford manifold (G, X) and an embedding $(f, u): (G_0, X_0) \hookrightarrow (G, X)$.

Construction A. Let (T_0, x_0) be a point with complex multiplication in (G_0, X_0) . Choose a CM-field (L, F) splitting T_0 . We construct an embedding

$$(f, u): (G_0, X_0) \hookrightarrow (G, X)$$

as above, setting $G' = G_{0F}$, $G = R_{F/\mathbf{Q}}G'$. We set $T' = T_{0F}$, $T = R_{F/\mathbf{Q}}T'$, $x = u(x_0)$; then (T, x) is a point with complex multiplication in (G, X) . Next, we construct a family of Mumford submanifolds of type A_1 in (G, X) passing

through x . For any root $\alpha \in R := R(G_{0L}, T_{0L}) \subset \text{Hom}(T_L, \mathbf{G}_{mL})$, let H'_α denote the centralizer of $(\ker \alpha)^0$ in G' (note that T' is F -anisotropic, hence $(\ker \alpha)^0$ is defined over F). Set $G'_\alpha = (H'_\alpha)^{\ker}$; then $G'_L \simeq \text{SL}_{2L}$. Set $G_\alpha = R_{F/\mathbf{Q}} G'_\alpha$. Let I denote the set of $\alpha \in R$ such that $G_\alpha(\mathbf{R})$ is noncompact. For $\alpha \in I$ set $X_\alpha = G_\alpha(\mathbf{R}) \cdot x$. Then (G_α, X_α) is a Mumford submanifold of type A_1 in (G, X) . For tangent spaces we have $T_x(X) = \bigoplus_{\alpha \in I} T_x(X_\alpha)$. Set $T_\alpha = T \cap G_\alpha$. Let T_α^{ad} denote the image of T in G^{ad} and let $p_\alpha: T^{\text{ad}} \rightarrow T_\alpha^{\text{ad}}$ be the canonical epimorphism. One can prove that $\bigcap_{\alpha \in I} \ker p_\alpha = 1$ (it follows from the inclusion $I + I \supset R$).

Using Construction A we shall deduce $C_1(G_0, X_0, x_0)$ from $K(G, X)$ and $C_1(G_\alpha, X_\alpha, x)$ ($\alpha \in I$). By Proposition 6(1) it suffices to prove $C_1(G, X, x)$. Suppose $\alpha \in I$ and set $b_\alpha = p_\alpha(t) \in \mathcal{P}(T_\alpha^{\text{ad}})$ (with the notation of the Lemma in §3). Then b_α determines a subgroup ${}^\tau G_\alpha \subset {}^\tau G$ such that $\varphi(G_\alpha(\mathbf{A}^f)) = {}^\tau G_\alpha(\mathbf{A}^f)$. From $\psi(\tau[e, x]) = [e, x']$ it follows that $\psi(\tau \text{Sh}(G_\alpha, X_\alpha)) = \text{Sh}({}^\tau G_\alpha, {}^\tau X_\alpha)$, where ${}^\tau X_\alpha = {}^\tau G_\alpha(\mathbf{R}) \cdot x'$. The statement $C_1(G_\alpha, X_\alpha, x)$ determines t_α uniquely; it implies that $t_\alpha = p_\alpha(b(T^{\text{ad}}, x, \tau))$. Since $\bigcap_{\alpha \in I} \ker p_\alpha = 1$, then $t = b(T^{\text{ad}}, x, \tau)$ is what was to be proved.

Now assuming that K and C_1 are proved, we can similarly deduce C_2 for (G_0, X_0) from C_2 for Mumford manifolds of type A_1 . Let x_0 and x_0^* be two CM-points of X_0 . If there is a Mumford submanifold passing through x_0 and x_0^* , then by Proposition 6(2a) $C_2(G_0, X_0, x_0, x_0^*)$ holds. Even if we could connect x_0 with x_0^* by a path consisting of Mumford submanifolds of type A_1 , Conjecture $C_2(G_0, X_0, x_0, x_0)$ would be valid due to Proposition 6(2c). Moreover, it suffices to draw this path in a larger Mumford manifold $(G, X) \supset (G_0, X_0)$. Therefore C_2 follows from

THEOREM B. *Let (G_0, X_0) be a simple Mumford manifold and let x_0, x_0^* be two CM-points of X_0 . Then there is an embedding $(f, u): (G_0, X_0) \hookrightarrow (G, X)$ and a finite sequence x_1, \dots, x_n of CM-points of X such that each of the pairs $(x := u(x_0), x_1), (x_1, x_2), \dots, (x_{n-1}, x_n)$ lies in a Mumford submanifold of type A_1 and that $x_n = \gamma \cdot u(x_0)$ for some $\gamma \in G(\mathbf{Q})$.*

It remains to be said about the proof of Conjecture K . It is deduced from Kazhdan's theorem [9] with the help of Margulis's results [10]. Some technical difficulties arise because of our inadequate knowledge of the structure of $G(\mathbf{Q})$. We overcome these difficulties using Piatetski-Shapiro's construction again; it enables us to consider instead of $G(\mathbf{Q})$ the group $G(\overline{\mathbf{Q}}_r)$, where $(\overline{\mathbf{Q}}_r)$ is the field of all the totally real algebraic numbers in \mathbf{R} .

Thus, Langlands's conjecture is proved for all the Mumford manifolds, and so is Shimura's conjecture, because by [14] Langlands's conjecture implies Shimura's conjecture. The proof outlined above follows the author's paper [3] (see also [2]). It is clear from the preceding discussion that the proof heavily depends on methods and results of a number of works, notably those of Mumford [15], Piatetski-Shapiro [16], Shimura [18], Deligne [4, 5, 6], Langlands [10], Milne and Shih [13, 14], Kazhdan [9], and Margulis [10]. Note that a similar proof

of Langlands's conjecture was obtained by Milne [12] who used some ideas and results of the author's unpublished notes containing a partial proof of Shimura's conjecture.

Our proof of Langlands's conjecture depends on the theory of complex multiplication of abelian varieties. We consider this dependence as a shortcoming of the proof. It is desirable to find a proof independent of the theory of abelian varieties. In this connection note that Faltings [7] proved the existence of canonical models over $\overline{\mathbf{Q}}$ of Shimura varieties, using only their intrinsic properties (especially rigidity). In a series of papers of Baily (see, e.g., [1]) a theory of arithmetic Hilbert modular functions is constructed independently of abelian varieties.

Since Shimura's conjecture is proved, we have an arithmetic theory of automorphic functions on Shimura varieties. It is, however, very desirable to have an arithmetic theory of automorphic *forms* (this is the second problem of Shimura [19]). This problem under some restrictions is solved by M. Harris [8].

Speaking about arithmetic automorphic functions we mean arithmeticity in the sense of Shimura. Shimura defined arithmetic automorphic functions in terms of their values at CM-points. There is another definition of arithmeticity in terms of coefficients of Fourier-Jacoby series. Harris [8] announced the coincidence of these two definitions in the cases to which they both apply.

REFERENCES

1. W. L. Baily, Jr., *Arithmetic Hilbert modular functions*. II, *Revista Mat. Iberoamer.* **1** (1985), 85–119.
2. M. V. Borovoi, *Langlands' conjecture on conjugation of Shimura varieties*, *Functional Anal. Appl.* **16** (1982), no. 4, 292–294.
3. —, *Langlands' conjecture concerning conjugation of connected Shimura varieties*, *Selecta Math. Soviet.* **3** (1983), no. 1, 3–39.
4. P. Deligne, *Travaux de Shimura*, *Séminaire Bourbaki*, 23ème année (1970/71), Exp. No. 389, *Lecture Notes in Math.*, Vol. 244, Springer, Berlin, 1971, pp. 123–165.
5. —, *Variétés de Shimura: interprétation modulaire et techniques de construction de modèles canoniques*, *Proc. Sympos. Pure Math.*, Vol. 33, Part 2, Amer. Math. Soc., Providence, R. I., 1979, pp. 247–290.
6. —, *Motifs et groupes de Tanigama*, *Lecture Notes in Math.*, Vol. 900, Springer, Berlin, 1982, pp. 261–279.
7. G. Faltings, *Arithmetic varieties and rigidity*, *Semin. Théor. Nombr.* (Paris, 1982/83), Boston, 1984, pp. 63–77.
8. M. Harris, *Arithmetic vector bundles and automorphic forms on Shimura varieties*. I, *Invent. Math.* **82** (1985), 151–189.
9. D. Kazhdan, *On arithmetic varieties*. II, *Israel J. Math.* **44** (1983), 139–159.
10. R. P. Langlands, *Automorphic representations, Shimura varieties and motives. Ein Märchen*, *Proc. Sympos. Pure Math.*, Vol. 33, Part 2, Amer. Math. Soc., Providence, R. I., 1979, pp. 205–246.
11. G. Margulis, *Discrete groups of motions of manifolds of nonpositive curvature*, *Amer. Math. Soc. Transl.* **109** (1977), 33–45.
12. J. S. Milne, *The action of an automorphism of \mathbf{C} on a Shimura variety and its special points*, *Progress in Math.*, Vol. 35, Birkhäuser, Boston, 1981, pp. 239–265.
13. J. S. Milne and K.-y. Shih, *Langlands' construction of the Taniyama group*, *Lecture Notes in Math.*, Vol. 900, Springer, Berlin, 1982, pp. 229–260.
14. —, *Conjugates of Shimura varieties*, *Lecture Notes in Math.*, Vol. 900, Springer, Berlin, 1982, pp. 280–356.

15. D. Mumford, *A note on Shimura's paper "Discontinuous groups and abelian varieties"*, Math. Ann. **181** (1969), 345–351.
16. I. I. Piatetski-Shapiro, *Interrelations between the Tate and Hodge conjectures for abelian varieties*, Math. USSR-Sb. **14** (1971), 615–625.
17. I. Satake, *Symplectic representations of algebraic groups satisfying a certain analyticity condition*, Acta Math. **117** (1967), 215–279.
18. G. Shimura, *On canonical models of arithmetic quotients of bounded symmetric domains*. I, Ann. of Math. **91** (1970), 144–22; II, **92** (1970), 528–549.
19. ———, *On some problems of algebraicity*, Proc. Internat. Congr. Math. (Helsinki, 1978), Vol. 1, Acad. Sci. Fennica, Helsinki, 1980, pp. 373–379.

1ST MOSCOW MEDICAL INSTITUTE, MOSCOW, USSR

Base Change for $\mathrm{GL}(n)$

L. CLOZEL

This article is a report on the results proved jointly with J. Arthur in [1]. We obtain the existence of a base change lift for automorphic forms on $\mathrm{GL}(n)$ under a solvable extension of number fields. For $\mathrm{GL}(2)$, such results had been obtained by Doi, Naganuma, Jacquet, Saito, Shintani, and Langlands [7, 14, 20, 21, 19b]. As noticed by Langlands for $\mathrm{GL}(2)$, base change may be used to shed new light on Artin's conjecture on the holomorphicity of L -functions. We do not know if solvable base change alone will yield new examples of the truth of Artin's conjecture. A result in that direction is proved in §3. We hope that the question will interest a specialist of finite groups.¹

1. Base change: Definitions and results.

1.1. Let G (or G_n if we want to specify the rank) denote the algebraic group $\mathrm{GL}(n)$ over \mathbf{Z} . Let K/k be a finite extension of number fields or of local fields of *characteristic zero*. In the global case, we denote by v a place (finite or infinite) of k , by w a place of K . For global k , let \mathbf{A}_k be the ring of adeles, \mathbf{A}_k^\times the group of ideles. Base change relates, in the local case, admissible irreducible representations [6] of $G(K)$ and $G(k)$; in the global case, automorphic representations [5] of $G(\mathbf{A}_K)$ and $G(\mathbf{A}_k)$. Let \mathcal{O}_k be the ring of integers of k .

In the global case, an automorphic representation can be written $\pi = \bigotimes_v \pi_v$, where π_v are local representations. For almost all finite v , π_v has a vector fixed by $G(\mathcal{O}_{k_v})$, and Hecke theory associates to it an element $t_{\pi,v}$ of $(\mathbf{C}^\times)^n / \mathfrak{S}_n$. We will often regard $t_{\pi,v}$ as a diagonal matrix and call it a *Hecke matrix*. We will denote by $\mathcal{A}(G/k)$ the set of automorphic (resp. admissible) irreducible representations of $G(\mathbf{A}_k)$ (resp. $G(k)$).

We want to restrict the class of automorphic representations considered. Let $n = n_1 + \cdots + n_r$ be a partition. Assume that π_i ($i = 1, \dots, r$) is a cuspidal, *unitary* representation of $G_{n_i}(\mathbf{A}_k)$. The group $M = G_{n_1} \times \cdots \times G_{n_r}$ is naturally identified to a Levi subgroup of G . We may extend the representation $\pi_1 \otimes \cdots \otimes \pi_r$

Partially supported by a Sloan Fellowship and National Science Foundation Grant DMS-8600003.

¹*Note added in proof:* Because of a theorem of E. C. Dade, base change does not yield new cases of the Artin conjecture. See footnote in §3.

of $M(\mathbf{A}_k)$ to the corresponding parabolic subgroup, and then unitarily induce to $G(\mathbf{A}_k)$. The resulting representation will be denoted by $\pi_1 \times \cdots \times \pi_r$. It is automorphic and irreducible. Such a representation of $G(\mathbf{A}_k)$ will be called *induced from cuspidal*. We denote by $I(G_n/k)$ the set of (equivalence classes of) such representations of $G_n(\mathbf{A}_k)$. If k is local, $I(G_n/k)$ denotes the set of tempered irreducible representations of $G_n(k)$.

1.2. Base change is most naturally described in terms of Langlands's "principle of functoriality" [19a, 4]. If G is a k -group, let ${}^L G = {}^L G^0 \rtimes \mathfrak{G}_k$ denote its L -group; here ${}^L G^0$ is a complex reductive group, on which $\mathfrak{G}_k = \text{Gal}(\bar{k}/k)$ acts by holomorphic automorphisms. We consider the following L -groups:

a. $G = \text{GL}(n)/k$. ${}^L G = \text{GL}(n, \mathbb{C}) \times \mathfrak{G}_k = {}^L G^0 \times \mathfrak{G}_k$.

b. Let H be the k -group obtained from $\text{GL}(n)/K$ by restriction of scalars. Then [4, §I.5]:

$${}^L H = ({}^L G^0 \times \cdots \times {}^L G^0) \rtimes \mathfrak{G}_k.$$

There is one copy of ${}^L G^0$ for each embedding of K into \bar{k} over k , and \mathfrak{G}_k acts by permuting the embeddings.

According to the principle of functoriality, any homomorphism between L -groups should be reflected by a correspondence between automorphic forms. We consider the following homomorphisms:

Restriction. Here ${}^L \rho_{k/K}: {}^L G \rightarrow {}^L H$ is given by the diagonal embedding: ${}^L G^0 \rightarrow {}^L H^0$. Conjecturally, one expects a mapping

$$\rho_{k/K}: \mathcal{A}(G/k) \rightarrow \mathcal{A}(G/K). \quad (?1.1)$$

(Note that canonically, $\mathcal{A}(H/k) = \mathcal{A}(G/K)!$)

Induction. Here ${}^L \iota_K^k: {}^L H \rightarrow {}^L G_{dn}$, $d = [K:k]$, is given by sending ${}^L H^0 \cong {}^L G_n^0 \times \cdots \times {}^L G_n^0$, as block-diagonal matrices, into ${}^L G_{dn}^0$. One extends it to an L -homomorphism by letting \mathfrak{G}_k act on ${}^L G_{dn}^0$ by permutation matrices in \mathfrak{S}_d (see [19d]). It should be mirrored by a mapping

$$\iota_K^k: \mathcal{A}(G_n/K) \rightarrow \mathcal{A}(G_{dn}/k). \quad (?1.2)$$

In the global case, the effect of $\rho_{k/K}$ and ι_K^k on Hecke matrices is explicitly described by the L -group. For $\text{GL}(n)$, because of the "strong multiplicity-one theorem" of Jacquet-Shalika [15b], this implies that ρ and ι are uniquely defined if we restrict ourselves to representations in $I(G_n/k)$.

EXAMPLE 1.1. Assume that K/k is cyclic, and consider ι_K^k for $n = 1$. In this case, ι_K^k sends Grössencharaktere χ on \mathbf{A}_K^\times to automorphic representations of $\text{GL}(d, \mathbf{A}_k)$. For $d = 2$, it is the mapping $\chi \mapsto \pi(\chi)$ of Hecke, Maass, Weil, Jacquet-Langlands [16, 21]. For general d , it has been obtained by Kazhdan and Flicker [17, 9] following the technique introduced by Labesse-Langlands [18].

1.3. In the global case, it is enlightening to interpret base change in terms of the conjectural Tannaka group $\mathcal{G}_{\Pi(k)}$ [19c]. The finite-dimensional, complex representations of degree n of this group should be in bijection with isobaric [19c, 15b] automorphic representations of $G_n(\mathbf{A}_k)$. (Representations in $I(G_n/k)$ are

isobaric.) If K/k is a Galois extension, one would expect an exact sequence

$$1 \rightarrow \mathcal{G}_{\Pi(K)} \rightarrow \mathcal{G}_{\Pi(k)} \rightarrow \mathfrak{G}_{K/k} \rightarrow 1 \quad (?1.3)$$

with $\mathfrak{G}_{K/k} = \text{Gal}(K/k)$. The factors $\rho_{k/K}$ and ι_K^k should respectively translate, for automorphic representations, the operations of restriction and induction of complex representations of the Tannaka groups. They may be called *automorphic restriction* and *induction*. The usual properties of these operations extend to them. In particular, for three extensions $K/L/k$:

$$\rho_{k/K} = \rho_{k/L} \circ \rho_{L/K}, \quad \iota_K^k = \iota_L^k \circ \iota_K^L. \quad (?1.4)$$

In the local case, this conjectural description remains appropriate, replacing $\mathcal{G}_{\Pi(k)}$ by the local Weil-Deligne group. Of course much is known in this case [9, 10, 12].

1.4. For k global, K/k cyclic of prime degree l , ρ and ι can be explicitly described by their effect on Hecke matrices at the unramified primes. Let ζ be a primitive l th root of unity.

$$\rho_{k/K}: \pi \mapsto \Pi, \quad \begin{cases} t_{\Pi,w} = t_{\pi,v} & (w \mid v, v \text{ split in } K), \\ t_{\Pi,w} = t_{\pi,v} & (w \mid v, v \text{ inert}). \end{cases} \quad (1.5)$$

$$\iota_K^k: \Pi \rightarrow \pi, \quad \begin{cases} t_{\pi,v} = t_{\Pi,w_1} \oplus \cdots \oplus t_{\Pi,w_l}, \\ \quad \quad \quad v \text{ split into } w_1, \dots, w_l, \\ t_{\pi,v} = t_{\Pi,w}^{1/l} \oplus \zeta t_{\Pi,w}^{1/l} \oplus \cdots \oplus \zeta^{l-1} t_{\Pi,w}^{1/l}, \\ \quad \quad \quad v \text{ inert, } w \mid v. \end{cases} \quad (1.6)$$

In the second line of (1.6), $t_{\Pi,w}^{1/l}$ denotes any choice of an l th root of $t_{\Pi,w}$. Note that $t_{\pi,v}$ is still unambiguously defined modulo the symmetric group on nl letters.

Note that formulas (?1.4), (1.5), and (1.6) completely specify ρ and ι for *solvable* extensions and representations in I . Thus the only problem is existence.

1.5. We now state our main result.

THEOREM 1.2. *Let K/k be a solvable extension of number fields.*

(i) *Assume $\pi \in I(G_n/k)$. Then there is a (unique) representation $\Pi \in I(G_n/K)$ such that $\Pi = \rho_{k/K}(\pi)$.*

(ii) *Assume $\Pi \in I(G_n/K)$. Then there is a (unique) representation $\pi \in I(G_{dn}/\rho)$ such that $\pi = \iota_K^k(\Pi)$.*

In the local case, we remark that ρ and ι should be compatible to (i) parabolic induction, sending representations of a Levi subgroup to representations of $G(F)$ ($F = k, K$); (ii) Local-global consistency: if K/k is a global extension and v a place of k inert in K , the effect of $\rho_{K/k}$ on the local component of an automorphic representation at v should be ρ_{K_v/k_v} ; the same applies to ι .

THEOREM 1.3. *Let K/k be a Galois extension of p -adic fields. Then there exist unique mappings $\rho_{k/K}: I(G_n/k) \rightarrow I(G_n/K)$ and $\iota_K^k: I(G_n/K) \rightarrow I(G_{dn}/k)$ compatible to the global functors and parabolic induction.*

2. Cyclic case. Because of (?1.4), the proofs restrict to the case of cyclic extensions. Assume now that K/k is *cyclic*. In that case, the results are more explicit; in particular, we characterize the image of ρ .

2.1. Local results. (K/k local non-Archimedean). Let σ be a generator of $\text{Gal}(K/k)$. Recall [20] that there is a norm map \mathcal{N} sending elements of $G(K)$ to conjugacy classes of $G(k)$. Two smooth, compactly supported functions ϕ on $G(K)$ and f on $G(k)$ are called *associated* if the twisted orbital integral of ϕ at $\delta \in G(K)$ is equal (for suitable normalizations) to the orbital integral of f at $\gamma = \mathcal{N}\delta$; γ is assumed regular.

Assume K/k unramified. For $F = K, k$, let \mathcal{H}_F be the Hecke algebra of functions on $G(F)$, bi-invariant by $G(\mathcal{O}_F)$. The morphism ${}^L\rho_{k/K}$ defines a homomorphism of algebras $b: \mathcal{H}_K \rightarrow \mathcal{H}_k$.

THEOREM 2.1. *If $\phi \in \mathcal{H}_K$, ϕ and $f = b\phi \in \mathcal{H}_k$ are associated.*

When ϕ, f are the *unit elements* of the Hecke algebras, Theorem 2.1 is due to Kottwitz [11]. Our proof relies crucially on his result.

The local results follow from Theorem 2.1 and the Deligne-Kazhdan trace formula. We obtain the following characterization of $\rho_{k/K}$. Let Π, π stand for representations in $I(G_n/K), I(G_n/k)$. Assume Π is σ -stable: $\Pi \cong \Pi \circ \sigma$. Recall that Π *lifts* π , in Shintani's sense, if the character of π , composed with \mathcal{N} , is equal to the twisted character of Π .

PROPOSITION 2.2. (i) *Given π , there is a unique σ -stable Π lifting π .*
(ii) *Conversely, if Π is σ -stable, there is (at least one) π lifted by Π .*

We set $\rho_{k/K}(\pi)$ equal to the unique Π given by (i). This completely determines the functor ρ by local conditions.

2.2. An identity of traces (K/k global). The global results are extracted from an identity of trace formulas. Let $G(k_\infty) = \prod_{v \text{ infinite}} G(k_v)$: it is a real Lie group. We denote by Z_k the center of its enveloping algebra. If Z_K is the analog for K , there is a natural homomorphism $N: Z_K \rightarrow Z_k$. We fix a character ν of Z_k ; let $\nu_K = \nu \circ N$.

For simplicity set $\mathbf{A} = \mathbf{A}_k$. Let M_0 be a maximal split k -torus in G . Let \mathcal{L} be the set of Levi subgroups $M \supseteq M_0$. For $M \in \mathcal{L}$, W_0^M is the Weyl group of M_0 in M . Let A_M be the split component of M , \mathfrak{a}_M its "real Lie algebra" [2], $W(\mathfrak{a}_M)$ the corresponding Weyl group, $M(\mathbf{A})^1$ the subgroup of $M(\mathbf{A})$ on which rational characters have absolute value 1. If $f \in C_c^\infty(G(\mathbf{A})^1)$, and π belongs to the unitary dual $\Pi(M(\mathbf{A})^1)$ of $M(\mathbf{A})^1$, let $\rho_\pi(\nu, f)$ denote, for $\nu \in i\mathfrak{a}_M^*$, the representation of $G(\mathbf{A})^1$ induced from the part of the discrete spectrum of $L^2(M(\mathbf{A})^1)$ that is isotypic of type π . Let $W(\mathfrak{a}_M)_{\text{reg}}$ be the set of $s \in W(\mathfrak{a}_M)$ whose fixed vectors are exactly \mathfrak{a}_G . We then set

$$T_{\text{dis}, \nu}(f) = \sum_{M \in \mathcal{L}} \sum_{s \in W(\mathfrak{a}_M)_{\text{reg}}} \sum_{\pi \in \Pi(M(\mathbf{A})^1)} \frac{|W_0^M|}{|W_0^G|} |\det(s-1)|^{-1} \cdot \text{trace}(M(s, 0) \rho_\pi(0, f)).$$

Here $M(s, 0)$ is an intertwining operator given by the theory of Eisenstein series; the sum runs only on representations having infinitesimal character ν .

The expression $T_{\mathrm{dis}, \nu}(f)$ represents the discrete part of the trace formula evaluated at f . There is an analogous formula for the twisted trace of $\phi \in C_c^\infty(G(\mathbf{A}_K)^1)$. Note that σ acts naturally on automorphic forms on $G(\mathbf{A}_K)$, and in the space of ρ_π for $\pi \in \Pi(M(\mathbf{A}_K)^1)$. We set

$$T_{\mathrm{dis}, \nu_K}^\sigma(\phi) = \sum_{M \in \mathcal{L}} \sum_{s \in W(\mathfrak{a}_M)_{\mathrm{reg}}} \sum_{\pi \in \Pi(M(\mathbf{A}_K)^1)} \frac{|W_0^M|}{|W_0^G|} |\det(s - 1)|^{-1} \cdot \mathrm{trace}(M(s, 0) \sigma \rho_\pi(0, \delta)).$$

The sum runs only on representations with infinitesimal character ν_K .

We say that ϕ, f are associated if their local components are, in the manner of §2.1.

THEOREM 2.3. *Assume ϕ, f are associated. Then*

$$T_{\mathrm{dis}, \nu}(f) = l T_{\mathrm{dis}, \nu_K}^\sigma(\phi).$$

We also obtain the similar identity of traces between the discrete part of the trace of f , and of the trace of a function $f^1 \in C_c^\infty(G^1(\mathbf{A})^1)$ associated to f , when G^1 is an inner form of G . The proof relies on Arthur's recent work on the invariant form of the "fine" expression of the trace formula.

2.3. Global lifting (K/k global). We denote by π (resp. Π) a representation of $I(G_n/k)$ (resp. $I(G_n/K)$). We say that Π lifts π if $\Pi_v = \bigotimes_{w|v} \Pi_w$ lifts π_v at all places v of k .

We will assume that $l = [K : k]$ is prime. Let η be a character of \mathbf{A}_k^\times associated by class field theory to K/k ; thus $\eta^l = 1$. If $\pi \in I(G_n/k)$, $\pi \otimes \eta$ denotes π twisted by η considered as a 1-dimensional character.

THEOREM 2.4. (i) *Assume π is cuspidal, $\pi \not\cong \pi \otimes \eta$. Then there is a unique σ -stable Π lifting π ; Π is cuspidal.*

(ii) *Assume π is cuspidal, $\pi \cong \pi \otimes \eta$. Then there is a cuspidal $\Pi_1 \in I(G_{n/l}, K)$, with $\Pi_1 \not\cong \Pi_1^\sigma$, such that $\Pi = \Pi_1 \times \Pi_1^\sigma \times \cdots \times \Pi_1^{\sigma^{l-1}}$ lifts π .*

(iii) *Assume Π is cuspidal, $\Pi \cong \Pi \circ \sigma$. Then there is a cuspidal π lifted by Π ; all such π are obtained from one of them by twisting by powers of η ; they satisfy $\pi \not\cong \pi \otimes \eta$.*

(iv) *Assume that $n = lm$. Let Π_1 be a cuspidal representation of $\mathrm{GL}(m, \mathbf{A}_k)$, $\Pi_1 \not\cong \Pi_1^\sigma$. Then $\Pi = \Pi_1 \times \Pi_1^\sigma \times \cdots \times \Pi_1^{\sigma^{l-1}}$ is σ -stable and lifts a unique cuspidal π ; $\pi \cong \pi \otimes \eta$.*

Theorem 2.4 follows from Theorem 2.3 and a systematic use of the results of Jacquet-Shalika [15a, b]. Parts (iii) and (iv) imply the existence of induction in Theorem 1.1.

3. Applications. Let K/k be a Galois extension of number fields, with Galois group \mathfrak{G} . If r is a complex representation of \mathfrak{G} , let $L(s, r)$ be the Artin L -function associated to r . It is a product $L(s, r) = \prod_v L_v(s, r)$ over the places of k .

If $\pi = \bigotimes \pi_v$ is an automorphic representation of $G(\mathbf{A}_k)$, let

$$L(s, \pi) = \prod_v L(s, \pi_v)$$

denote its standard L -function [13].

We will call “strong Artin conjecture” the following assertion. If r is irreducible, there should exist a cuspidal representation π of $\mathrm{GL}(n, \mathbf{A}_k)$ such that, for almost all v :

$$L_v(s, r) = L(s, \pi_v). \quad (3.1)$$

This conjecture is due to Langlands [19a]. It implies Artin’s conjecture that $L(s, r)$ is holomorphic if r is irreducible and nontrivial.

We will say that r is automorphic if, assuming r irreducible, a cuspidal π satisfying (3.1) for almost all v exists; $\pi = \pi(r)$ is then unique.

Assume \mathfrak{H} is a subgroup of \mathfrak{G} . We say that \mathfrak{H} is subnormal if there is a composition sequence (\mathfrak{H}_i) for \mathfrak{H} in \mathfrak{G} with cyclic quotients.

THEOREM 3.1. *Assume $\mathfrak{G} = \mathrm{Gal}(K/k)$ is solvable. Assume that r , an irreducible representation of \mathfrak{G} , belongs to the submodule of the Grothendieck group of characters of \mathfrak{G} spanned over \mathbb{Z} by characters of the form $\mathrm{ind}_{\mathfrak{H}}^{\mathfrak{G}} \chi$, \mathfrak{H} subnormal in \mathfrak{G} , χ an Abelian character.*

Then r is automorphic.

In particular, the strong Artin conjecture is true for such a representation. We do not know if nonmonomial representations of this type exist: if so, that would provide new examples of Artin’s conjecture.²

COROLLARY 3.2. *Assume \mathfrak{G} is nilpotent. Then the strong Artin conjecture is true for K/k .*

Theorem 3.1 results from automorphic induction and a theorem of [15b] allowing one to construct differences of automorphic representations.

The proof of these results has an interesting twist. Assume that π, τ are automorphic representations of $G_m(\mathbf{A}_k), G_n(\mathbf{A}_k)$ respectively. A fundamental problem in the theory of automorphic forms is to construct an automorphic representation Π of $G_{mn}(\mathbf{A}_k)$ such that, at almost all v , $t_{\Pi, v} = t_{\pi, v} \otimes t_{\tau, v}$. It is denoted by $\pi \boxtimes \tau$ in [19c].

Now assume \mathfrak{G} nilpotent, and let $\pi(r)$ be associated to an irreducible representation r of degree d . Let π be any cuspidal representation of $G_n(\mathbf{A}_k)$. Then $\pi(r) \boxtimes \pi$ exists. Indeed, since r is monomial, we may write $\pi(r) = \iota_L^k(\chi)$ where χ , a character of $\mathfrak{H} = \mathrm{Gal}(K/L)$, is seen as a character of \mathbf{A}_L^\times . Though $\mathfrak{H} \subset \mathfrak{G}$ is only subnormal, automorphic induction applies. The formula

$$\iota_L^k(\chi \boxtimes \rho_{k/L} \pi) = \iota_L^k \chi \boxtimes \pi, \quad (3.2)$$

which is just, for finite groups, the familiar formula about induction, restriction, and tensor product, then shows that $\pi(r) \boxtimes \pi$ exists.

² *Note added in proof:* Professor E. C. Dade has shown that all such representations are monomial. Therefore the Artin conjecture was already known to be true for them.

REFERENCES

1. J. Arthur and L. Clozel, *Base change for $GL(n)$* , Preprint.
2. J. Arthur, *Eisenstein series and the trace formula*, Proc. Sympos. Pure Math., vol. 33, part I, Amer. Math. Soc., Providence, R.I., 1979, pp. 253–275.
3. A. Borel and W. Casselman (Editors), *Automorphic forms, representations and L -functions*, Proc. Sympos. Pure Math., vol. 33, parts I–II, Amer. Math. Soc., Providence, R.I., 1979.
4. A. Borel, *Automorphic L -functions*, Proc. Sympos. Pure Math., vol. 33, part II, Amer. Math. Soc., Providence, R.I., 1979, pp. 27–62.
5. A. Borel and H. Jacquet, *Automorphic forms and automorphic representations*, Proc. Sympos. Pure Math., vol. 33, part I, Amer. Math. Soc., Providence, R.I., 1979, pp. 189–202.
6. P. Cartier, *Representations of p -adic groups*, Proc. Sympos. Pure Math., vol. 33, part I, Amer. Math. Soc., Providence, R. I., 1979, pp. 111–156.
7. K. Doi and H. Naganuma, *On the functional equation of certain Dirichlet series*, Invent. Math. **9** (1969), 1–14.
8. Y. Flicker, *On twisted lifting*, Trans. Amer. Math. Soc. **290** (1985), 161–178.
9. G. Henniart, *On the local Langlands conjecture for $GL(p)$* , Preprint.
10. —, *On the local Langlands conjecture for $GL(n)$: the cyclic case*, Ann. of Math. (2) **123** (1986), 146–203.
11. R. Kottwitz, *Orbital integrals on GL_3* , Amer. J. Math. **102** (1980), 327–384.
12. P. Kutzko and A. Moy, *On the local Langlands conjecture in prime dimension*, Ann. of Math. (2) **121** (1985), 495–517.
13. R. Godement and H. Jacquet, *Zeta functions of simple algebras*, Lecture Notes in Math., vol. 260, Springer-Verlag, 1972.
14. H. Jacquet, *Automorphic forms on $GL(2)$. II*, Lecture Notes in Math., vol. 278, Springer-Verlag, 1972.
- 15a. H. Jacquet and J. Shalika, *On Euler products and the classification of automorphic representations. I*, Amer. J. Math **103** (1981), 499–558.
- 15b. —, *On Euler products and the classification of automorphic representations. II*, Amer. J. Math **103** (1981), 777–815.
16. H. Jacquet and R. P. Langlands, *Automorphic forms on $GL(2)$* , Lecture Notes in Math., vol. 114, Springer-Verlag, 1970.
17. D. Kazhdan, *On lifting*, Lie Group Representations. II (Herb, editor), Lecture Notes in Math., vol. 1041, Springer-Verlag, 1984.
18. J. P. Labesse and R. P. Langlands, *L -indistinguishability for $SL(2)$* , Canad. J. Math. **31** (1979), 726–785.
- 19a. R. P. Langlands, *Problems in the theory of automorphic forms*, Lectures in Modern Analysis and Applications, Lecture Notes in Math., vol. 170, Springer-Verlag, 1970, pp. 18–86.
- 19b. —, *Base change for $GL(2)$* , Princeton Univ. Press, Princeton, N.J., 1980.
- 19c. —, *Automorphic representations, Shimura varieties, and motives*, Proc. Sympos. Pure Math., vol. 33, part II, Amer. Math. Soc., Providence, R.I., 1979, pp. 205–246.
- 19d. —, *Les débuts d'une formule des traces stable*, Publ. Math. Univ. Paris VII, no. 13, Université de Paris VII, U.E.R. de Mathématiques, Paris, 1983.
20. H. Saito, *Automorphic forms and algebraic extensions of number fields*, Lectures in Math., no. 8, Kinokuniya Book-Store, Tokyo, 1975.
- 21a. T. Shintani, *On liftings of holomorphic automorphic forms*, Proc. U.S.-Japan Seminar on Number Theory, Ann Arbor, 1975.
- 21b. —, *On liftings of holomorphic automorphic forms*, Proc. Sympos. Pure Math., vol. 33, part II, Amer. Math. Soc., Providence, R.I., 1979, pp. 97–110.
22. A. Weil, *Dirichlet series and automorphic forms*, Lecture Notes in Math., vol. 189, Springer-Verlag, 1971.

UNIVERSITY OF MICHIGAN, ANN ARBOR, MICHIGAN 48109, USA

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE, PARIS, FRANCE

Quantum Groups

V. G. DRINFEL'D

This is a report on recent works on Hopf algebras (or quantum groups, which is more or less the same) motivated by the quantum inverse scattering method (QISM), a method for constructing and studying integrable quantum systems, which was developed mostly by L. D. Faddeev and his collaborators. Most of the definitions, constructions, examples, and theorems in this paper are inspired by the QISM. Nevertheless I will begin with these definitions, constructions, etc. and then explain their relation to the QISM. Thus I reverse the history of the subject, hoping to make its logic clearer.

1. What is a quantum group? Recall that both in classical and in quantum mechanics there are two basic concepts: state and observable. In classical mechanics states are points of a manifold M and observables are functions on M . In the quantum case states are 1-dimensional subspaces of a Hilbert space H and observables are operators in H (we forget the self-adjointness condition). The relation between classical and quantum mechanics is easier to understand in terms of observables. Both in classical and in quantum mechanics observables form an associative algebra which is commutative in the classical case and noncommutative in the quantum case. So quantization is something like replacing commutative algebras by noncommutative ones.

Now let us consider elements of a group G as states and functions on G as observables. The notion of group is usually defined in terms of states. To quantize it one has to translate it first into the language of observables. This translation is well known, but let us recall it nevertheless. Consider the algebra $A = \text{Fun}(G)$ consisting of functions on G which are supposed to be smooth if G is a Lie group, regular if G is an algebraic group, and so on. A is a commutative associative unital algebra. Clearly $\text{Fun}(G \times G) = A \otimes A$ if one understands the sign " \otimes " in the appropriate sense (e.g. if G is a Lie group then \otimes should be understood as the topological tensor product). So the group operation considered as a mapping $f: G \times G \rightarrow G$ induces an algebra homomorphism $\Delta: A \rightarrow A \otimes A$ called "comultiplication." To formulate the associativity property of the group

operation in terms of Δ one expresses it as the commutativity of the diagram

$$\begin{array}{ccccc} & & f \times \text{id} & \rightarrow & G \times G & \xrightarrow{f} & G \\ G \times G \times G & & & & & & \\ & & \text{id} \times f & \rightarrow & G \times G & \xrightarrow{f} & G \end{array}$$

and then the functor $X \mapsto \text{Fun}(X)$ is applied. The result is the coassociativity property of Δ , i.e., the commutativity of the diagram

$$\begin{array}{ccccc} & & \Delta & \rightarrow & A \otimes A & \xrightarrow{\text{id} \otimes \Delta} & A \otimes A \otimes A \\ A & & & & & & \\ & & \Delta & \rightarrow & A \otimes A & \xrightarrow{\Delta \otimes \text{id}} & A \otimes A \otimes A \end{array}$$

The translation of the notions of the group unit e and the inversion mapping $g \mapsto g^{-1}$ is quite similar. We consider the mappings $\varepsilon: A \rightarrow k$, $S: A \rightarrow A$ given by $\varphi \mapsto \varphi(e)$ and $\varphi(g) \mapsto \varphi(g^{-1})$ respectively (here k is the ground field, i.e. $k = \mathbf{R}$ if G is a real Lie group, $k = \mathbf{C}$ if G is a complex Lie group, etc). Then we translate the identities $g \cdot e = e \cdot g = g$, $g \cdot g^{-1} = g^{-1} \cdot g = e$ into the language of commutative diagrams and apply the functor $X \mapsto \text{Fun}(X)$. As a result we obtain the commutative diagrams

$$\begin{array}{ccc} A & \xrightarrow{\text{id}} & A \\ \Delta \downarrow & & \parallel \\ A \otimes A & \xrightarrow{\text{id} \otimes \varepsilon} & A \otimes k \end{array} \quad \begin{array}{ccc} A & \xrightarrow{\text{id}} & A \\ \Delta \downarrow & & \parallel \\ A \otimes A & \xrightarrow{\varepsilon \otimes \text{id}} & k \otimes A \end{array} \quad (1)$$

$$\left. \begin{array}{ccccc} A & \xrightarrow{\Delta} & A \otimes A & \xrightarrow{\text{id} \otimes S} & A \otimes A & \xrightarrow{m} & A \\ & \searrow \varepsilon & & & & \nearrow i & \\ & & k & & & & \\ A & \xrightarrow{\Delta} & A \otimes A & \xrightarrow{S \otimes \text{id}} & A \otimes A & \xrightarrow{m} & A \\ & \searrow \varepsilon & & & & \nearrow i & \\ & & k & & & & \end{array} \right\} \quad (2)$$

Here m is the multiplication (i.e. $m(a \otimes b) = ab$) and $i(c) = c \cdot 1_A$. The commutativity of (1) (resp. (2)) is expressed by the words “ ε is the counit” (resp. “ S is the antipode”). The properties of $(A, m, \Delta, i, \varepsilon, S)$ listed above mean that A is a commutative Hopf algebra.

Now there is a general principle: the functor $X \mapsto \text{Fun}(X)$ from the category of “spaces” to the category of commutative associative unital algebras (perhaps, with some additional structures or properties) is an antiequivalence. This principle becomes a theorem if “space” is understood as “affine scheme,” or if “space”

is understood as "compact topological space" and "algebra" is understood as " C^* -algebra." From the above principle it follows that the category of groups is antiequivalent to the category of commutative Hopf algebras.

Now let us define the category of quantum spaces to be dual to the category of (not necessarily commutative) associative unital algebras. Denote by $\text{Spec } A$ (spectrum of A) the quantum space corresponding to an algebra A . Finally define a quantum group to be the spectrum of a (not necessarily commutative) Hopf algebra. So the notions of Hopf algebra and quantum group are in fact equivalent, but the second one has some geometric flavor.

Let me make some general remarks on the definition of Hopf algebra. First of all, it is known that for given A, m, Δ the counit $\varepsilon: A \rightarrow k$ is unique and is a homomorphism. The antipode $S: A \rightarrow A$ is also unique and it is an antihomomorphism (with respect to both multiplication and comultiplication). Secondly, in the noncommutative case one also requires the existence of the "skew antipode" $S': A \rightarrow A$ which is in fact the antipode for the opposite multiplication and the same comultiplication. S' is also the antipode for the opposite comultiplication and the same multiplication. It is known that $SS' = S'S = \text{id}$. In the commutative or cocommutative case $S' = S$, but in general $S' \neq S$ and $S^2 \neq \text{id}$. The proofs of these statements can be found in standard texts on Hopf algebras [1–3]. An informal discussion of some aspects of the notion of Hopf algebra can be found in [4]. The reader should keep in mind that our usage of the term "Hopf algebra" is not generally accepted: some people do not require the existence of the unit, the counit and the antipode (so their Hopf algebras correspond to quantum semigroups).

It is important that a quantum group is not a group, nor even a group object in the category of quantum spaces. This is because for noncommutative algebras the tensor product is not a coproduct in the sense of category theory.

Now the question arises whether there exist natural examples of noncommutative Hopf algebras. An easy way of constructing such an algebra is to consider A^* , A being a commutative but not cocommutative algebra (recall that the dual space of a Hopf algebra A has a Hopf algebra structure, the multiplication mapping $A^* \otimes A^* \rightarrow A^*$ being induced by the comultiplication of A and the comultiplication of A^* being induced by the multiplication of A). In this way more or less all cocommutative noncommutative Hopf algebras are obtained. The words "more or less" are due to the fact that $V^{**} \neq V$ for a general vector space V . But at the heuristic level $V^{**} = V$ and therefore any cocommutative Hopf algebra is of the form $(\text{Fun}(G))^*$ for some group G . Clearly $(\text{Fun}(G))^*$ is commutative iff G is commutative. Note that $(\text{Fun}(G))^*$ is nothing but the group algebra of G . An important class of cocommutative Hopf algebras is formed by universal enveloping algebras (the comultiplication $U\mathfrak{g} \rightarrow U\mathfrak{g} \otimes U\mathfrak{g}$ is defined by the formula $\Delta(x) = x \otimes 1 + 1 \otimes x$, $x \in \mathfrak{g}$). If \mathfrak{g} is the Lie algebra of a Lie group G then $U\mathfrak{g}$ may be considered as the subalgebra of $(C^\infty(G))^*$ consisting of distributions $\varphi \in C_0^{-\infty}(G)$ such that $\text{Supp } \varphi \subset \{e\}$. One can identify $(U\mathfrak{g})^*$ with the completion of the local ring of $e \in G$ or with $\text{Fun}(\hat{G})$ where \hat{G} is the

formal group corresponding to \mathfrak{g} (the second realization of $(U\mathfrak{g})^*$ makes sense even if G does not exist, which may well happen if $\dim \mathfrak{g} = \infty$).

The most interesting and mysterious Hopf algebras are those which are neither commutative nor cocommutative. Though Hopf algebras were intensively studied both by “pure” algebraists [2–12] and by specialists in von Neumann algebras [13–26], I believe that most of the examples of noncommutative noncocommutative Hopf algebras invented independently of integrable quantum system theory are counterexamples rather than “natural” examples (however there are remarkable exceptions, e.g. [12], [16], and [2, pp. 89–90]). We are going to discuss a general method for constructing noncommutative noncocommutative Hopf algebras, which was proposed in [27, 28] under the influence of the QISM. This method is based on the concept of quantization. It can be considered as a realization of the ideas of G. I. Kac and V. G. Palyutkin (see the end of [16]).

2. Quantization. Roughly speaking, a quantization of a commutative associative algebra A_0 over k is a (not necessarily commutative) deformation of A_0 depending on a parameter \hbar (Planck’s constant), i.e. an associative algebra A over $k[[\hbar]]$ such that $A/\hbar A = A_0$ and A is a topologically free $k[[\hbar]]$ -module. Given A , we can define a new operation on A_0 (the Poisson bracket) by the formula

$$\{a \bmod \hbar, b \bmod \hbar\} = \frac{[a, b]}{\hbar} \bmod \hbar. \quad (3)$$

Thus A_0 becomes a Poisson algebra (i.e. a Lie algebra with respect to $\{, \}$ and a commutative associative algebra with respect to multiplication, these two structures being compatible in the following sense:

$$\{a, bc\} = \{a, b\}c + b\{a, c\}.$$

Now we shall slightly change our point of view on quantization.

DEFINITION. A *quantization* of a Poisson algebra A_0 is an associative algebra deformation A of A_0 over $k[[\hbar]]$ such that the Poisson bracket on A_0 defined by (3) is equal to the bracket given a priori.

Of course, this approach to quantization is as old as quantum mechanics. It was explained to mathematicians by F. A. Berezin, J. Vey, A. Lichnerowicz, M. Flato, D. Sternheimer, and others.

We shall need a Hopf algebra version of the above definition. In this case A_0 is a Poisson-Hopf algebra (i.e. a Hopf algebra structure and a Poisson algebra structure on A_0 are given such that the multiplication is the same for both structures and the comultiplication $A_0 \rightarrow A_0 \otimes A_0$ is a Poisson algebra homomorphism, the Poisson bracket on $A_0 \otimes A_0$ being defined by $\{a \otimes b, c \otimes d\} = ac \otimes \{b, d\} + \{a, c\} \otimes bd$) and A is a Hopf algebra deformation of A_0 . We shall also use the dual notion of quantization of co-Poisson-Hopf algebras (a co-Poisson-Hopf algebra is a cocommutative Hopf algebra B with a Poisson cobracket $B \rightarrow B \otimes B$ compatible with the Hopf algebra structure).

We discuss the structure of Poisson-Hopf algebras and co-Poisson-Hopf algebras in §§3 and 4. Then we consider the quantization problem.

3. Poisson groups and Lie bialgebras. A *Poisson group* is a group G with a Poisson bracket on $\text{Fun}(G)$ which makes $\text{Fun}(G)$ a Poisson-Hopf algebra. In other words the Poisson bracket must be compatible with the group operation, which means that the mapping $\mu: G \times G \rightarrow G$, $\mu(g_1, g_2) = g_1 g_2$, must be a Poisson mapping in the sense of [33], i.e. $\mu^*: \text{Fun}(G) \rightarrow \text{Fun}(G \times G)$ must be a Lie algebra homomorphism. Specifying the meaning of the word "group" and the symbol $\text{Fun}(G)$, we obtain the notions of Poisson-Lie group, Poisson formal group, Poisson algebraic group, etc. According to our general principles the notions of Poisson group and Poisson-Hopf algebra are equivalent.

There exists a very simple description of Poisson-Lie groups in terms of *Lie bialgebras*.

DEFINITION. A *Lie bialgebra* is a vector space \mathfrak{g} with a Lie algebra structure and a Lie coalgebra structure, these structures being compatible in the following sense: the cocommutator mapping $\mathfrak{g} \rightarrow \mathfrak{g} \otimes \mathfrak{g}$ must be a 1-cocycle (\mathfrak{g} acts on $\mathfrak{g} \otimes \mathfrak{g}$ by means of the adjoint representation).

If G is a Poisson-Lie group then $\mathfrak{g} = \text{Lie}(G)$ has a Lie bialgebra structure. To define it write the Poisson bracket on $C^\infty(G)$ as

$$\{\varphi, \psi\} = \eta^{\mu\nu} \partial_\mu \varphi \cdot \partial_\nu \psi, \quad \varphi, \psi \in C^\infty(G), \quad (4)$$

where $\{\partial_\mu\}$ is a basis of right-invariant vector fields on G . The compatibility of the bracket with the group operation means that the function $\eta: G \rightarrow \mathfrak{g} \otimes \mathfrak{g}$ corresponding to $\eta^{\mu\nu}$ is a 1-cocycle. The 1-cocycle $f: \mathfrak{g} \rightarrow \mathfrak{g} \otimes \mathfrak{g}$ corresponding to η defines a Lie bialgebra structure on \mathfrak{g} (the Jacobi identity for $f^*: \mathfrak{g}^* \otimes \mathfrak{g}^* \rightarrow \mathfrak{g}^*$ holds because f^* is the infinitesimal part of the bracket (4)).

THEOREM 1. *The category of connected and simply-connected Poisson-Lie groups is equivalent to the category of finite dimensional Lie bialgebras.*

The analogue of Theorem 1 for Poisson formal groups over a field of characteristic 0 can be proved in the following way. The algebra of functions on the formal group corresponding to \mathfrak{g} is nothing but $(U\mathfrak{g})^*$. A Poisson-Hopf structure on $(U\mathfrak{g})^*$ is equivalent to a co-Poisson-Hopf structure on $U\mathfrak{g}$. So it suffices to prove the following easy theorem.

THEOREM 2. *Let $\delta: U\mathfrak{g} \rightarrow U\mathfrak{g} \otimes U\mathfrak{g}$ be a Poisson cobracket which makes $U\mathfrak{g}$ a co-Poisson-Hopf algebra (the Hopf structure on $U\mathfrak{g}$ is usual). Then $\delta(\mathfrak{g}) \subset \mathfrak{g} \otimes \mathfrak{g}$ and $(\mathfrak{g}, \delta|_{\mathfrak{g}})$ is a Lie bialgebra. Thus we obtain a one-to-one correspondence between co-Poisson-Hopf structures on $U\mathfrak{g}$ inducing the usual Hopf structure and Lie bialgebra structures on \mathfrak{g} inducing the given Lie algebra structure.*

Now let us discuss the notion of Lie bialgebra. First of all there is a one-to-one correspondence between Lie bialgebras and *Manin triples*. A Manin triple $(\mathfrak{p}, \mathfrak{p}_1, \mathfrak{p}_2)$ consists of a Lie algebra \mathfrak{p} with a nondegenerate invariant scalar product on it and isotropic Lie subalgebras $\mathfrak{p}_1, \mathfrak{p}_2$ such that \mathfrak{p} is the direct sum of \mathfrak{p}_1 and \mathfrak{p}_2 as a vector space. The correspondence mentioned above is constructed in the following way: if $(\mathfrak{p}, \mathfrak{p}_1, \mathfrak{p}_2)$ is a Manin triple then we put $\mathfrak{g} = \mathfrak{p}_1$ and define the

cocommutator $\mathfrak{g} \rightarrow \mathfrak{g} \otimes \mathfrak{g}$ to be dual to the commutator mapping $\mathfrak{p}_2 \otimes \mathfrak{p}_2 \rightarrow \mathfrak{p}_2$ (note that \mathfrak{p}_2 is naturally isomorphic to \mathfrak{g}^*). Conversely, if a Lie bialgebra \mathfrak{g} is given we put $\mathfrak{p} = \mathfrak{g} \oplus \mathfrak{g}^*$, $\mathfrak{p}_1 = \mathfrak{g}$, $\mathfrak{p}_2 = \mathfrak{g}^*$ and define the commutator $[x, l]$ for $x \in \mathfrak{g}$, $l \in \mathfrak{g}^*$ so that the natural scalar product on \mathfrak{p} should be invariant. Note that if $(\mathfrak{p}, \mathfrak{p}_1, \mathfrak{p}_2)$ is a Manin triple then so is $(\mathfrak{p}, \mathfrak{p}_2, \mathfrak{p}_1)$. Therefore the notion of Lie bialgebra is self-dual.

Here are some examples of Lie bialgebras. Examples 3.2–3.4 are important for the inverse scattering method.

EXAMPLE 3.1. If $\dim \mathfrak{g} = 2$ then any linear mappings $\bigwedge^2 \mathfrak{g} \rightarrow \mathfrak{g}$ and $\mathfrak{g} \rightarrow \bigwedge^2 \mathfrak{g}$ define a Lie bialgebra structure on \mathfrak{g} . A 2-dimensional Lie bialgebra is called nondegenerate if the composition $\bigwedge^2 \mathfrak{g} \rightarrow \mathfrak{g} \rightarrow \bigwedge^2 \mathfrak{g}$ is nonzero. In this case there exists a basis $\{x_1, x_2\}$ of \mathfrak{g} such that $[x_1, x_2] = \alpha x_2$ and the cocommutator is given by $x_1 \mapsto 0, x_2 \mapsto \beta x_2 \wedge x_1$. Here $\alpha\beta \neq 0$ and $\alpha\beta$ does not depend on the choice of x_1, x_2 .

EXAMPLE 3.2. Let \mathfrak{g} be a Kac-Moody algebra (in the sense of [55]) with a fixed invariant scalar product $\langle \cdot, \cdot \rangle$, \mathfrak{h} the Cartan subalgebra, $\mathfrak{b}_\pm \supset \mathfrak{h}$ the Borel subalgebras. Put $\mathfrak{p} = \mathfrak{g} \times \mathfrak{g}$, $\mathfrak{p}_1 = \{(x, y) \in \mathfrak{g} \times \mathfrak{g} | x = y\} \approx \mathfrak{g}$, $\mathfrak{p}_2 = \{(x, y) \in \mathfrak{b}_- \times \mathfrak{b}_+ | x_{\mathfrak{h}} + y_{\mathfrak{h}} = 0\}$. Define the scalar product of $(x_1, y_1) \in \mathfrak{p}$ and $(x_2, y_2) \in \mathfrak{p}$ to equal $\langle x_1, x_2 \rangle - \langle y_1, y_2 \rangle$. Since $(\mathfrak{p}, \mathfrak{p}_1, \mathfrak{p}_2)$ is a Manin triple, \mathfrak{g} has a Lie bialgebra structure. The cocommutator $\varphi: \mathfrak{g} \rightarrow \bigwedge^2 \mathfrak{g}$ can be described explicitly in terms of the canonical generators X_i^\pm, H_i (here $X_i^\pm \in [\mathfrak{b}_\pm, \mathfrak{b}_\pm]$ and H_i is the image of the simple root $\alpha_i \in \mathfrak{h}^*$ under the isomorphism $\mathfrak{h}^* \rightarrow \mathfrak{h}$): $\varphi(H_i) = 0$, $\varphi(X_i^\pm) = \frac{1}{2} X_i^\pm \wedge H_i$. Note that \mathfrak{b}_+ and \mathfrak{b}_- are subbialgebras of \mathfrak{g} . If $\mathfrak{g} = \mathfrak{sl}(2)$ then \mathfrak{b}_+ and \mathfrak{b}_- are of the type described in Example 3.1. The Manin triple corresponding to \mathfrak{b}_+ is $(\mathfrak{g} \times \mathfrak{h}, \text{Im} \psi_+, \text{Im} \psi_-)$ where $\psi_\pm: \mathfrak{b}_\pm \hookrightarrow \mathfrak{g} \times \mathfrak{h}$ is defined by $\psi_\pm(a) = (a, \pm a_{\mathfrak{h}})$ and the scalar product on $\mathfrak{g} \times \mathfrak{h}$ equals $\langle \cdot, \cdot \rangle_{\mathfrak{g}} - \langle \cdot, \cdot \rangle_{\mathfrak{h}}$.

EXAMPLE 3.3. Fix a simple Lie algebra \mathfrak{a} ($\dim \mathfrak{a} < \infty$) and an invariant scalar product on it. Set $\mathfrak{p} = \mathfrak{a}((u^{-1}))$, $\mathfrak{p}_1 = \mathfrak{a}[u]$, $\mathfrak{p}_2 = u^{-1}\mathfrak{a}[[u^{-1}]]$ and define the scalar product on \mathfrak{p} by $(f, g) = \text{res}_{u=\infty}(f(u), g(u)) du$. The Manin triple $(\mathfrak{p}, \mathfrak{p}_1, \mathfrak{p}_2)$ defines a Lie bialgebra structure on $\mathfrak{g} = \mathfrak{a}[u]$. The cocommutator in \mathfrak{g} is given by the formula

$$a(u) \mapsto [a(u) \otimes 1 + 1 \otimes a(v), r(u, v)]. \quad (5)$$

Here $a \in \mathfrak{a}[u]$, $\mathfrak{g} \otimes \mathfrak{g}$ is identified with $(\mathfrak{a} \otimes \mathfrak{a})[u, v]$ and $r(u, v) = t/(u - v)$ where t is the element of $\mathfrak{a} \otimes \mathfrak{a}$ corresponding to our scalar product. The right-hand side of (5) has no pole at $u = v$ because t is invariant.

EXAMPLE 3.4 (cf. [29–31]). Fix a nonsingular irreducible projective algebraic curve X over \mathbb{C} . Denote by E (resp. A, O_x) the field of rational functions on X (resp. the adèle ring, the completion of the local ring of X at x). Fix an absolutely simple Lie algebra G over E ($\dim G < \infty$) and a rational differential ω on X , $\omega \neq 0$. Define an invariant \mathbb{C} -valued scalar product on $G \otimes_E A$ by

$$(u, v) = \sum_{x \in X} \text{res}_x \{ \omega \cdot \text{Tr} \rho(u_x) \rho(v_x) \}$$

where $\rho: G \rightarrow \mathfrak{gl}(n, E)$ is an exact representation, $u = (u_x)_{x \in X}$, $v = (v_x)_{x \in X}$. Then G is a maximal isotropic subspace of $G \otimes_E A$. If there exists an open isotropic \mathbf{C} -subalgebra $\Lambda \subset G \otimes_E A$ such that $G \otimes_E A = G \oplus \Lambda$ then we obtain a Lie bialgebra structure on G . To every subset $S \subset X$, $S \neq \emptyset$, there corresponds a subbialgebra $G_S = \{a \in G \mid \text{the image of } a \text{ in } G \otimes A^S \text{ belongs to the image of } \Lambda \text{ in } G \otimes A^S\}$, where A^S is the ring of adèles without x -components, $x \in S$. The bialgebra of Example 3.3 is, in fact, G_S , where

$$X = \mathbf{P}^1, \quad \omega = d\lambda, \quad S = \{\infty\}, \quad \mathfrak{g} = \mathfrak{a} \otimes E, \quad \Lambda = \mathfrak{a} \otimes \left(m_\infty \times \prod_{x \in X \setminus S} O_x \right),$$

and m_∞ is the maximal ideal of O_∞ . If \mathfrak{g} is an affine Kac-Moody algebra with the bialgebra structure of Example 3.2 then $\mathfrak{g}' \stackrel{\text{def}}{=} [\mathfrak{g}, \mathfrak{g}] / (\text{the center of } \mathfrak{g})$ is, in fact, G_S for $X = \mathbf{P}^1$, $\omega = \lambda^{-1} d\lambda$, $S = \{0, \infty\}$, $G = \mathfrak{g}' \otimes_{\mathbf{C}[\lambda, \lambda^{-1}]} \mathbf{C}(\lambda)$ (\mathfrak{g}' has a natural structure of a $\mathbf{C}[\lambda, \lambda^{-1}]$ -algebra),

$$\Lambda = \mathfrak{c} \times \prod_{x \in X \setminus S} (\mathfrak{g}' \otimes_{\mathbf{C}[\lambda, \lambda^{-1}]} O_x),$$

where $\mathfrak{c} \subset (\mathfrak{g}' \otimes_{\mathbf{C}[\lambda, \lambda^{-1}]} \mathbf{C}((\lambda^{-1}))) \times (\mathfrak{g}' \otimes_{\mathbf{C}[\lambda, \lambda^{-1}]} \mathbf{C}((\lambda)))$ is the closure of

$$\mathfrak{p}'_2 = \{(x, y)\} \in \mathfrak{g}' \times \mathfrak{g}' \mid x \in \mathfrak{b}'_-, y \in \mathfrak{b}'_+, x_\eta + y_\eta = 0\}.$$

4. Classical Yang-Baxter equation. Let \mathfrak{g} be a Lie algebra and suppose that $\varphi: \mathfrak{g} \rightarrow \wedge^2 \mathfrak{g}$ is the coboundary of $r \in \wedge^2 \mathfrak{g}$. It can be shown that (\mathfrak{g}, φ) is a Lie bialgebra iff

$$[r^{12}, r^{13}] + [r^{12}, r^{23}] + [r^{13}, r^{23}] \text{ is } \mathfrak{g}\text{-invariant.} \quad (6)$$

Here, for instance,

$$[r^{12}, r^{13}] \stackrel{\text{def}}{=} \sum_{i,j} [a_i, a_j] \otimes b_i \otimes b_j$$

where $r = \sum_i a_i \otimes b_i$ (one may imagine that $r^{12} = \sum_i a_i \otimes b_i \otimes 1 \in (U\mathfrak{g})^{\otimes 3}$, $r^{13} = \sum_i a_i \otimes 1 \otimes b_i \in (U\mathfrak{g})^{\otimes 3}$, $r^{23} = \sum_i 1 \otimes a_i \otimes b_i \in (U\mathfrak{g})^{\otimes 3}$). In particular, if

$$[r^{12}, r^{13}] + [r^{12}, r^{23}] + [r^{13}, r^{23}] = 0, \quad (7)$$

then (\mathfrak{g}, φ) is a Lie bialgebra. Equation (7) is just the classical Yang-Baxter equation (CYBE), or the classical triangle equation. There is an important case which is intermediate between (6) and (7). Suppose that $r \in \mathfrak{g} \otimes \mathfrak{g}$ satisfies (7) but the skew-symmetry condition $r^{12} + r^{21} = 0$ is replaced by " $r^{12} + r^{21}$ is \mathfrak{g} -invariant" (then the coboundary $\varphi = \partial r$ is skew-symmetric). In this situation (\mathfrak{g}, φ) is also a Lie bialgebra. If $r^{12} + r^{21} = P$, $\rho = r - \frac{1}{2}P$ then $\rho \in \wedge^2 \mathfrak{g}$, $\varphi = \partial \rho$ and $[\rho^{12}, \rho^{13}] + [\rho^{12}, \rho^{23}] + [\rho^{13}, \rho^{23}] = \frac{1}{4}[P^{12}, P^{23}]$.

DEFINITION. A *coboundary Lie bialgebra* is a pair (\mathfrak{g}, r) , where \mathfrak{g} is a Lie bialgebra, $r \in \wedge^2 \mathfrak{g}$, ∂r = the cocommutator of \mathfrak{g} . A coboundary Lie bialgebra (\mathfrak{g}, r) is said to be *triangular* if r satisfies (7). A *quasitriangular Lie bialgebra* is

a pair (\mathfrak{g}, r) , where \mathfrak{g} is a Lie bialgebra, $r \in \mathfrak{g} \otimes \mathfrak{g}$, ∂r = the cocommutator of \mathfrak{g} , and r satisfies (7).

It was noticed in [32, 27] that the expression $[r^{12}, r^{13}] + [r^{12}, r^{23}] + [r^{13}, r^{23}]$, $r \in \bigwedge^2 \mathfrak{g}$, is equal to $\frac{1}{2}\{r, r\}$ where $\{, \}$ is the bilinear operation on $\bigwedge^* \mathfrak{g}$ such that $\{a, b\} = [a, b]$ for $a, b \in \mathfrak{g}$ and

$$\begin{aligned}\{a, b\} &= -(-1)^{(k+1)(l+1)}\{b, a\}, \\ \{a, b \wedge c\} &= \{a, b\} \wedge c + (-1)^{(k+1)l}b \wedge \{a, c\}\end{aligned}$$

for $a \in \bigwedge^k \mathfrak{g}$, $b \in \bigwedge^l \mathfrak{g}$, $c \in \bigwedge^* \mathfrak{g}$ ($\bigwedge^* \mathfrak{g}$ is a Poisson superalgebra with an odd Poisson bracket; if elements of $\bigwedge^* \mathfrak{g}$ are considered as left-invariant polyvector fields on the Lie group corresponding to \mathfrak{g} then the operation $\{, \}$ in $\bigwedge^* \mathfrak{g}$ is the Schouten bracket [34]).

The Poisson bracket on the Poisson-Lie group G corresponding to a coboundary bialgebra (\mathfrak{g}, r) is of the form

$$\{\varphi, \psi\} = r^{\mu\nu}(\partial'_\mu \varphi \cdot \partial'_\nu \psi - \partial_\mu \varphi \cdot \partial_\nu \psi), \quad (8)$$

where ∂_μ (respectively ∂'_μ) are the right-invariant (respectively left-invariant) vector fields on G corresponding to a basis of \mathfrak{g} and $r^{\mu\nu}$ are the coordinates of r . The following property of the bracket (8) is crucial for the Hamiltonian version of the classical inverse scattering method (a systematic exposition of this method is given in [35]; see also [36–38]):

$$\varphi, \psi \in C^\infty(G), \quad \varphi(gyg^{-1}) = \varphi(y), \quad \psi(gyg^{-1}) = \psi(y) \Rightarrow \{\varphi, \psi\} = 0. \quad (9)$$

Let us reconsider Examples 3.1–3.4. The cocommutator of a nondegenerate 2-dimensional bialgebra is not a coboundary. Now let us consider the cocommutator (5) on $\mathfrak{g} = \mathfrak{a}[u]$. First suppose that $r(u, v)$ is an $\mathfrak{a} \otimes \mathfrak{a}$ -valued polynomial. Then we can consider r as an element of $\mathfrak{g} \otimes \mathfrak{g}$ and the cocommutator (5) as the coboundary of r . The CYBE and the condition $r \in \bigwedge^2 \mathfrak{g}$ take the form

$$\begin{aligned}[r^{12}(u_1, u_2), r^{13}(u_1, u_3)] + [r^{12}(u_1, u_2), r^{23}(u_2, u_3)] \\ + [r^{13}(u_1, u_3), r^{23}(u_2, u_3)] = 0,\end{aligned} \quad (10)$$

$$r^{12}(u_1, u_2) + r^{21}(u_2, u_1) = 0. \quad (11)$$

So every polynomial solution of (10), (11) defines a triangular bialgebra structure on \mathfrak{g} . In Example 3.3, $r(u, v)$ is a nonpolynomial solution of (10), (11). Therefore the corresponding bialgebra is “pseudotriangular” (nevertheless (9) holds).

The bialgebra \mathfrak{g} of Example 3.2 has a quasitriangular structure if $\dim \mathfrak{g} < \infty$: consider the element $t \in \mathfrak{g} \otimes \mathfrak{g}$ corresponding to the scalar product on \mathfrak{g} , represent t as $t_{+-} + t_0 + t_{-+}$, $t_{+-} \in [\mathfrak{b}_+, \mathfrak{b}_+] \otimes [\mathfrak{b}_-, \mathfrak{b}_-]$, $t_{-+} \in [\mathfrak{b}_-, \mathfrak{b}_-] \otimes [\mathfrak{b}_+, \mathfrak{b}_+]$, $t_0 \in \mathfrak{h} \otimes \mathfrak{h}$, and set $r = t_{+-} + \frac{1}{2}t_0$. If $\dim \mathfrak{g} = \infty$ then r does not belong to the algebraic tensor product $\mathfrak{g} \otimes \mathfrak{g}$. Therefore (\mathfrak{g}, r) is “pseudoquasitriangular.” Let us examine more attentively the non-twisted affine case (the twisted affine case is similar). In this case $\mathfrak{g}' = [\mathfrak{g}, \mathfrak{g}]/(\text{the center of } \mathfrak{g}) = \mathfrak{a}[\lambda, \lambda^{-1}]$ for some simple algebra \mathfrak{a} , $\dim \mathfrak{a} < \infty$. Denote by \bar{r} the analogue of r for \mathfrak{g}' . Consider \bar{r} as a power series in $\lambda = \lambda \otimes 1$ and

$\mu = 1 \otimes \lambda$ with coefficients in $\mathfrak{a} \otimes \mathfrak{a}$. Then $\bar{r}(\lambda, \mu)$ is the Laurent decomposition of a rational function of λ/μ having a single pole at $\lambda/\mu = 1$ with residue $t_{\mathfrak{a}}$ ($t_{\mathfrak{a}} \in \mathfrak{a} \otimes \mathfrak{a}$ corresponds to the scalar product on \mathfrak{a}). This rational function satisfies (10) and (11). Moral [39, §3]: if a solution $r(u, v)$ of (10), (11) has a pole at $u = v$ then the corresponding Lie bialgebra is "pseudoquasitriangular" rather than "pseudotriangular"; in other words $r^{12}(u_1, u_2) + r^{21}(u_2, u_1)$ is $t_{\mathfrak{a}} \cdot \delta(u_1 - u_2)$ rather than 0.

I. V. Cherednik proved [31] that the bialgebras of Example 4 also correspond to solutions $r(u, v)$ of (10), (11) having a pole at $u = v$. Moreover, it follows from [40–42] that there is a one-to-one correspondence between "nondegenerate" solutions of (10), (11) up to an equivalence relation (such solutions always have a pole at $u = v$) and quadruples (X, ω, G, Λ) of the type described in Example 3.4. Besides, a classification of nondegenerate solutions of (10), (11) is given in [40, 41]. In terms of (X, ω, G, Λ) the results of [40, 41] can be stated in the following way.

(1) For Λ to exist ω should have no zeros and therefore there are three possibilities: (a) X is an elliptic curve and ω is regular, (b) $X = \mathbf{P}^1$, $\omega = \lambda^{-1}d\lambda$, (c) $X = \mathbf{P}^1$, $\omega = d\lambda$. The corresponding solutions $r(u, v)$ are elliptic, trigonometric or rational functions of $u - v$.

(2) In case (a) for Λ to exist G has to be isomorphic to $\mathfrak{sl}(n)$, and for every n there are finitely many choices of Λ all of which are listed. In case (b) G is of the type described at the end of Example 3.4, and all the choices of Λ are listed (they are similar to the Λ described at the end of Example 3.4). In case (c) little is known.

Some important ideas and results of M. A. Semenov-Tian-Shansky concerning Poisson-Lie groups and the CYBE are in [39, 43].

5. Some historical remarks. The Poisson bracket (8) and the CYBE were introduced by E. K. Sklyanin [44, 37] as the classical limits of the corresponding quantum objects. The compatibility of the bracket (8) with the group operation was formulated by him almost explicitly. The abstract notion of Poisson-Lie group appeared later [27]. The classification of the solutions of the CYBE was based on vast experimental material due to many authors (see the Appendix of [45]). In particular, the solution $r(u, v) = t/(u - v)$ was found in [45, 46].

6. Quantization of Lie bialgebras (examples). By definition, a quantization of a Lie bialgebra \mathfrak{g} is a quantization of $U\mathfrak{g}$ in the sense of §2, where $U\mathfrak{g}$ is considered as a co-Poisson-Hopf algebra according to Theorem 2. If A is a quantization of \mathfrak{g} we call g the *classical limit* of A .

EXAMPLE 6.1 (cf. [12], [2, pp. 89-90]). Consider the $\mathbf{C}[[\hbar]]$ -algebra A generated (in the \hbar -adic sense, i.e. as an algebra complete in the \hbar -adic topology) by x_1, x_2 with the defining relation $[x_1, x_2] = \alpha x_2$, $\alpha \in \mathbf{C}^*$. Define $\Delta: A \rightarrow A \otimes A$ by $\Delta(x_1) = x_1 \otimes 1 + 1 \otimes x_1$, $\Delta(x_2) = x_2 \otimes \exp(\frac{1}{2}\hbar\beta x_1) + \exp(-\frac{1}{2}\hbar\beta x_1) \otimes x_2$, $\beta \in \mathbf{C}^*$. Then A is a quantization of the bialgebra of Example 3.1. From

the results of §9 it follows that the quantization is unique up to substitutions $h \mapsto h + \sum_{i=2}^{\infty} c_i h^i, c_i \in \mathbb{C}$.

EXAMPLE 6.2. Let $\mathfrak{g}, \mathfrak{h}, \alpha_i, \langle \cdot, \cdot \rangle, H_i$ denote the same objects as in Example 3.2. Consider the $\mathbb{C}[[h]]$ -algebra $U_h \mathfrak{g}$ generated (in the h -adic sense) by \mathfrak{h} and X_i^+, X_i^- with the defining relations

$$[a_1, a_2] = 0 \quad \text{for } a_1, a_2 \in \mathfrak{h}, \quad [a, X_i^{\pm}] = \pm \alpha_i(a) X_i^{\pm} \quad \text{for } a \in \mathfrak{h},$$

and also

$$[X_i^+, X_j^-] = 2\delta_{ij} h^{-1} \text{sh}(hH_i/2),$$

$$\begin{aligned} i \neq j, \quad n = 1 - A_{ij}, \quad q = \exp \frac{h}{2} \langle H_i, H_i \rangle \Rightarrow \\ \Rightarrow \sum_{k=0}^n (-1)^k \binom{n}{k}_q q^{-k(n-k)/2} (X_i^{\pm})^k X_j^{\pm} (X_i^{\pm})^{n-k} = 0. \end{aligned}$$

Here (A_{ij}) is the Cartan matrix and $\binom{n}{k}_q$ is the Gauss polynomial [47, §3.3], i.e., $\binom{n}{k}_q = (q^n - 1)(q^{n-1} - 1) \cdots (q^{n-k+1} - 1) / (q^k - 1)(q^{k-1} - 1) \cdots (q - 1)$. It can be shown that $U_h \mathfrak{g}$ is a topologically free $\mathbb{C}[[h]]$ -module and that there exists a homomorphism $\Delta: U_h \mathfrak{g} \rightarrow U_h \mathfrak{g} \otimes U_h \mathfrak{g}$ such that

$$\Delta(X_i^{\pm}) = X_i^{\pm} \otimes \exp(hH_i/4) + \exp(-hH_i/4) \otimes X_i^{\pm},$$

$\Delta(a) = a \otimes 1 + 1 \otimes a$ for $a \in \mathfrak{h}$. $(U_h \mathfrak{g}, \Delta)$ is a quantization of \mathfrak{g} . It can be shown using the results of §9 that $U_h \mathfrak{g}$ is the unique (up to substitutions $h \mapsto h + \sum_{i=2}^{\infty} c_i h^i, c_i \in \mathbb{C}$) quantization A of \mathfrak{g} for which there exist a cocommutative Hopf subalgebra $C \subset A$ and a mapping $\theta: A \rightarrow A$ such that the mapping $C/hC \rightarrow U\mathfrak{g}$ is injective and its image is equal to $U\mathfrak{h}$, $\theta^2 = \text{id}$, $\theta(C) = C$, θ is an algebra automorphism and a coalgebra antiautomorphism, $\theta \bmod h$ is the Cartan involution (if $A = U_h \mathfrak{g}$ put $C = U\mathfrak{h}[[h]]$, $\theta(X_i^{\pm}) = -X_i^{\mp}$, $\theta|_{\mathfrak{h}} = -\text{id}$). So it is natural to call $U_h \mathfrak{g}$ the quantized Kac-Moody algebra corresponding to \mathfrak{g} . $U_h \mathfrak{sl}(2)$ was introduced by Kulish-Reshetikhin [48] and E. K. Sklyanin [49]. For general \mathfrak{g} the algebra $U_h \mathfrak{g}$ was introduced by M. Jimbo [50, 51] and the author [28]. All these works were motivated by the QISM. For \mathfrak{g} affine or finite dimensional $U_h \mathfrak{g}$ is closely related to trigonometric solutions of the quantum Yang-Baxter equation (QYBE); see [50–52] and §13.

EXAMPLE 6.3 [28, 54]. Let \mathfrak{a} and \mathfrak{g} denote the same objects as in Example 3.3. Since \mathfrak{g} is graded ($\deg au^k = k$ for $a \in \mathfrak{a}$) and the cocommutator has degree -1 , it is natural to look for a quantization Q which is a graded Hopf algebra over $\mathbb{C}[[h]]$, where $\mathbb{C}[[h]]$ is graded so that $\deg h = 1$. It can be shown using the results of §9 that Q exists and is unique. As an h -adic algebra, Q is generated by the Lie algebra \mathfrak{a} ($\deg a = 0$ for $a \in \mathfrak{a}$) and elements $J(a)$ of degree 1, $a \in \mathfrak{a}$, with the defining relations

$$J(\lambda a + \mu b) = \lambda J(a) + \mu J(b), \quad J([a, b]) = [a, J(b)] \quad \text{for } a, b \in \mathfrak{a}, \lambda, \mu \in \mathbb{C}, \quad (12)$$

$$\begin{aligned}\sum_i [a_i, b_i] &= 0 \Rightarrow \sum_i [J(a_i), J(b_i)] \\ &= \frac{h^2}{12} \sum_i ([a_i, I_\alpha], [b_i, I_\beta], I_\gamma) \cdot \{I_\alpha, I_\beta, I_\gamma\}, \quad a_i, b_i \in \mathfrak{a},\end{aligned}\quad (13)$$

$$\begin{aligned}\sum_i [[a_i, b_i], c_i] &= 0 \Rightarrow \sum_i [[J(a_i), J(b_i)], J(c_i)] \\ &= \frac{h^2}{4} \sum_i \sum_{\alpha, \beta, \gamma} f(a_i, b_i, c_i, I_\alpha, I_\beta, I_\gamma) \{I_\alpha, I_\beta, J(I_\gamma)\}, \quad a_i, b_i, c_i \in \mathfrak{a},\end{aligned}\quad (14)$$

where $\{I_\alpha\}$ is an orthonormal basis of \mathfrak{a} ,

$$\begin{aligned}f(a, b, c, x, y, z) &= \text{Alt Sym}_{\substack{a, b \\ a, c}}([a, [b, x]], [[c, y], z]), \\ \{x_1, x_2, x_3\} &= \frac{1}{6} \sum_{i \neq j \neq k} x_i x_j x_k.\end{aligned}$$

$\Delta: Q \rightarrow Q \otimes Q$ is defined by the formulas

$$\begin{aligned}\Delta(a) &= a \otimes 1 + 1 \otimes a, \\ \Delta(J(a)) &= J(a) \otimes 1 + 1 \otimes J(a) + \frac{h}{2} [a \otimes 1, t] \quad a \in \mathfrak{a},\end{aligned}\quad (15)$$

where $t = \sum_\alpha I_\alpha \otimes I_\alpha$. Note that the classical limit of $J(a)$ is $au \in \mathfrak{g}$. So the formula (12) and the left-hand sides of (13), (14) are quite natural. (15) is also natural because $[a \otimes 1, t]$ is the image of au under the cocommutator mapping. The terrific right-hand sides of (13), (14) are derived from (12), (15). Note also that if $\mathfrak{a} = \mathfrak{sl}(2)$ then (13) is superfluous and if $\mathfrak{a} \neq \mathfrak{sl}(2)$ then (14) is superfluous. The Hopf algebra over \mathbb{C} obtained from Q by setting $h = 1$ is denoted by $Y(\mathfrak{a})$ and called the Yangian of \mathfrak{a} . $Y(\mathfrak{a})$ is related to rational solutions of the QYBE (see §12), the simplest of which was found by C. N. Yang [53].

Fortunately Q has another system of generators $\xi_{ik}^\pm, \xi_{il}^\pm, \varphi_{ik}$, $i \in \Gamma, k = 0, 1, 2, \dots$, $\deg \xi_{ik}^\pm = \deg \varphi_{ik} = k$, where Γ is the set of simple roots of \mathfrak{a} . The classical limits of ξ_{ik}^\pm and φ_{ik} are equal to $X_i^\pm u^k$ and $H_i u^k$, where X_i^\pm and H_i are the generators of \mathfrak{a} used in Example 3.2. Here is the complete set of relations between ξ_{ik}^\pm and φ_{ik} :

$$\begin{aligned}[\varphi_{ik}, \varphi_{jl}] &= 0, \quad [\varphi_{i0}, \xi_{jl}^\pm] = \pm 2B_{ij} \xi_{jl}^\pm, \quad [\xi_{ik}^\pm, \xi_{jl}^\pm] = \delta_{ij} \varphi_{i, k+l}, \\ [\varphi_{i, k+1}, \xi_{jl}^\pm] - [\varphi_{ik}, \xi_{j, l+1}^\pm] &= \pm h B_{ij} (\varphi_{ik} \xi_{jl}^\pm + \xi_{jl}^\pm \varphi_{ik}) \\ [\xi_{i, k+1}^\pm, \xi_{jl}^\pm] - [\xi_{ik}^\pm, \xi_{j, l+1}^\pm] &= \pm h B_{ij} (\xi_{ik}^\pm \xi_{jl}^\pm + \xi_{jl}^\pm \xi_{ik}^\pm), \\ i \neq j, n = 1 - A_{ij} \Rightarrow \text{Sym}_{k_1, \dots, k_n} [\xi_{ik_1}^\pm, [\xi_{ik_2}^\pm, \dots [\xi_{ik_n}^\pm, \xi_{jl}^\pm], \dots]] &= 0,\end{aligned}$$

where (A_{ij}) is the Cartan matrix of \mathfrak{a} , $B_{ij} = \frac{1}{2}(H_i, H_j)$. The connection between the two presentations of Q is described in [54].

A similar realization exists for quantized affine Kac-Moody algebras [54]. For instance, if \mathfrak{g} is untwisted, i.e. $[\mathfrak{g}, \mathfrak{g}]/(\text{the center of } \mathfrak{g}) = \mathfrak{a}[\lambda, \lambda^{-1}]$, then $U_h \mathfrak{g}$ has the

following presentation. Generators: $c, D, \xi_{ik}^{\pm}, \kappa_{ik}$, where $i \in \Gamma$, $k \in \mathbf{Z}$. Defining relations:

$$\begin{aligned} [c, \kappa_{ik}] &= [c, \xi_{ik}^{\pm}] = [c, D] = 0, & [D, \kappa_{ik}] &= k\kappa_{ik}, & [D, \xi_{ik}^{\pm}] &= k\xi_{ik}^{\pm}, \\ [\kappa_{ik}, \kappa_{jl}] &= 4\delta_{k,-l}k^{-1}h^{-2}\text{sh}(khB_{ij})\text{sh}(khc/2), \\ [\kappa_{ik}, \xi_{jl}^{\pm}] &= \pm 2(kh)^{-1}\text{sh}(khB_{ij})\exp(\mp|k| \cdot hc/4)\xi_{j,k+l}^{\pm}, \\ \xi_{i,k+1}^{\pm}\xi_{jl}^{\pm} - e^{\pm hB_{ij}}\xi_{jl}^{\pm}\xi_{i,k+1}^{\pm} &= e^{\pm hB_{ij}}\xi_{ik}^{\pm}\xi_{j,l+1}^{\pm} - \xi_{j,l+1}^{\pm}\xi_{ik}^{\pm}, \\ [\xi_{ik}^+, \xi_{jl}^-] &= \delta_{ij}h^{-1}\{\psi_{i,k+l}e^{hc(k-l)/4} - \varphi_{i,k+l}e^{hc(l-k)/4}\}, \\ i \neq j, n = 1 - A_{ij}, q = \exp \frac{h}{2}(H_i, H_i) &\Rightarrow \\ \Rightarrow \text{Sym}_{k_1, \dots, k_n} \sum_{r=0}^n (-1)^r \binom{n}{r}_q q^{-r(n-r)/2} \xi_{ik_1}^{\pm} \cdots \xi_{ik_r}^{\pm} \xi_{jl}^{\pm} \xi_{ik_{r+1}}^{\pm} \cdots \xi_{ik_n}^{\pm} &= 0. \end{aligned}$$

Here φ_{ip}, ψ_{ip} are defined from the relations

$$\begin{aligned} \sum_p \varphi_{ip} u^p &= \exp h \left(\frac{1}{2} \kappa_{i0} + \sum_{p < 0} \kappa_{ip} u^p \right), \\ \sum_p \psi_{ip} u^p &= \exp h \left(\frac{1}{2} \kappa_{i0} + \sum_{p > 0} \kappa_{ip} u^p \right). \end{aligned}$$

The connection between the two presentations of $U_h \mathfrak{g}$ is described in [54]. Note that $U_h \mathfrak{g}'$ (i.e. $U_h \mathfrak{g}$ “without D ” and with the auxiliary relation $c = 0$) degenerates into $Y(\mathfrak{a})$: if A is the subalgebra of $U_h \mathfrak{g}' \otimes_{\mathbf{C}[[h]]} \mathbf{C}((h))$ generated by $U_h \mathfrak{g}'$ and $h^{-1} \cdot \text{Ker } f$, where f is the composition $U_h \mathfrak{g}' \rightarrow U \mathfrak{g}' = U(\mathfrak{a}[\lambda, \lambda^{-1}]) \xrightarrow{\lambda=1} U \mathfrak{a}$, then $A/hA = Y(\mathfrak{a})$.

The problem of describing Δ in terms of ξ_{ik}^{\pm} is not solved for Yangians nor for quantized affine algebras.

7. Quantized universal enveloping algebras and h -formal groups. A *quantized universal enveloping algebra* (QUE-algebra) is a Hopf algebra A over $\mathbf{C}[[h]]$ such that A is a topologically free $\mathbf{C}[[h]]$ -module and A/hA is a universal enveloping algebra. In other words, a QUE-algebra is a quantization of some Lie bialgebra. It is easy to show that a cocommutative QUE-algebra A is the h -adic completion of $U \mathfrak{g}$ for some Lie algebra \mathfrak{g} over $\mathbf{C}[[h]]$ which is a topologically free $\mathbf{C}[[h]]$ -module (in fact, $\mathfrak{g} = \{a \in A \mid \Delta(a) = a \otimes 1 + 1 \otimes a\}$).

Another important class of Hopf algebras over $\mathbf{C}[[h]]$ consists of *quantized formal series Hopf algebras* (QFSH-algebras). A QFSH-algebra is a topological Hopf algebra B over $\mathbf{C}[[h]]$ such that (1) as a topological $\mathbf{C}[[h]]$ -module B is isomorphic to $\mathbf{C}[[h]]^I$ for some set I , (2) B/hB is isomorphic to $\mathbf{C}[[u_1, u_2, \dots]]$ as a topological algebra.

An *h -formal group* is the spectrum of a QFSH-algebra. To understand h -formal groups in down-to-earth terms let us fix a “coordinate system,” i.e. elements $x_i \in B$ such that (1) B/hB is the ring of formal power series in $\bar{x}_i = x_i \bmod h$, and (2) $\varepsilon(x_i) = 0$, where $\varepsilon: B \rightarrow \mathbf{C}[[h]]$ is the counit. $\Delta(x_i)$ can be expressed as $F_i(x_1, x_2, \dots; y_1, y_2, \dots; h)$, where $y_i = 1 \otimes x_i \in B \otimes B$ and x_i is

identified with $x_i \otimes 1 \in B \otimes B$. So an \hbar -formal group is defined by a "quantum group law" $F(x, y, \hbar)$ and "commutation relations," i.e., formulas expressing $\hbar^{-1}[x_i, x_j]$ as a formal series in x_i and \hbar .

Now let us discuss the relation between QUE-algebras and QFSH-algebras. First of all, the dual of a QUE-algebra is a QFSH-algebra and vice versa (in both cases the dual should be understood in an appropriate sense). For instance, the dual of the QUE algebra of Example 6.1 is generated as an algebra by ξ_1 and ξ_2 where

$$\xi_1(f(x_1, x_2)) = \frac{\partial}{\partial x_1} f(x_1, 0)|_{x_1=0}, \quad \xi_2 \left(\sum_i \varphi_i(x_1) x_2^i \right) = \varphi_1 \left(\frac{\alpha}{2} \right).$$

The commutation relation between ξ_1 and ξ_2 is of the form $[\xi_1, \xi_2] = -\hbar\beta\xi_2$ and the comultiplication is given by $\Delta(\xi_1) = \xi_1 \otimes 1 + 1 \otimes \xi_1$, $\Delta(\xi_2) = \xi_2 \otimes \exp(-\frac{\alpha}{2}\xi_1) + \exp(\frac{\alpha}{2}\xi_1) \otimes \xi_2$. The dual of $U_{\hbar}\mathfrak{sl}(n)$ also has an explicit description. Consider the representation $\rho: U_{\hbar}\mathfrak{sl}(n) \rightarrow \text{Mat}(n, \mathbb{C}[[\hbar]])$ such that $\rho(H_i) = e_{ii} - e_{i+1, i+1}$, $\rho(X_i^+) = e_{i, i+1}$, $\rho(X_i^-) = 2\hbar^{-1} \text{sh}(\hbar/2) e_{i+1, i}$ where e_{ij} are the matrix units (the scalar product on $\mathfrak{sl}(n)$ is supposed to equal $\text{Tr}(XY)$). The matrix elements ρ_{ij} , $1 \leq i, j \leq n$, belong to $(U_{\hbar}\mathfrak{sl}(n))^*$. It is easy to show that

$$\rho_{ij}\rho_{il} = e^{-\hbar/2} \rho_{il}\rho_{ij} \quad \text{if } j < l, \quad (16)$$

$$\rho_{ij}\rho_{kj} = e^{-\hbar/2} \rho_{kj}\rho_{ij} \quad \text{if } i < k, \quad (17)$$

$$\rho_{il}\rho_{kj} = \rho_{kj}\rho_{il} \quad \text{if } i < k, \quad l > j, \quad (18)$$

$$[\rho_{il}, \rho_{kj}] = (e^{\hbar/2} - e^{-\hbar/2}) \rho_{ij}\rho_{kl} \quad \text{if } i > k, \quad l > j, \quad (19)$$

$$\sum_{i_1, \dots, i_n} \rho_{1i_1} \rho_{2i_2} \cdots \rho_{ni_n} \cdot (-e^{-\hbar/2})^{l(i_1, \dots, i_n)} = 1,$$

where $l(i_1, \dots, i_n)$ is the number of inversions in the permutation (i_1, \dots, i_n) . In fact $(U_{\hbar}\mathfrak{sl}(n))^*$ is the quotient of the ring of noncommutative power series in $\tilde{\rho}_{ij} = \rho_{ij} - \delta_{ij}$ by the ideal generated by the above relations. The comultiplication is usual: $\Delta(\rho_{ij}) = \sum_k \rho_{ik} \otimes \rho_{kj}$. It is natural to call $\text{Spec}(U_{\hbar}\mathfrak{sl}(n))^*$ the quantized formal $\text{SL}(n)$. One can also consider the subalgebra of $(U_{\hbar}\mathfrak{sl}(n))^*$ consisting of the matrix elements of finite-dimensional representations of $U_{\hbar}\mathfrak{sl}(n)$, i.e. of the polynomials in ρ_{ij} . The spectrum of this subalgebra can be considered as a quantization of the algebraic group $\text{SL}(n)$.

Somewhat unexpectedly the category of QFSH-algebras finitely generated as topological algebras is *equivalent* (not only antiequivalent) to the category of QUE-algebras whose classical limit is a finite dimensional Lie bialgebra (the finiteness conditions are not very essential here). The following example can help to understand this equivalence. Consider the QUE-algebra of Example 6.1, i.e. $[x_1, x_2] = \alpha x_2$, $\Delta(x_1) = x_1 \otimes 1 + 1 \otimes x_1$, $\Delta(x_2) = x_2 \otimes \exp(\frac{1}{2}\hbar\beta x_1) + \exp(-\frac{1}{2}\hbar\beta x_1) \otimes x_2$. Put $y_i = \hbar x_i$. Then $[y_1, y_2] = \alpha\hbar y_2$, $\Delta(y_1) = y_1 \otimes 1 + 1 \otimes y_1$, $\Delta(y_2) = y_2 \otimes \exp(\frac{1}{2}\beta y_1) + \exp(-\frac{1}{2}\beta y_1) \otimes y_2$, i.e. we have obtained a QFSH-algebra. Here is the general construction. If A is a QUE-algebra we put $\delta_1(a) = a - \varepsilon(a) \cdot 1$, $\delta_2(a) = \Delta(a) - a \otimes 1 - 1 \otimes a + \varepsilon(a) \cdot (1 \otimes 1)$, etc. (ε is the counit).

Then $\{a \in A \mid \delta_n(a) \in h^n A^{\otimes n} \text{ for any } n\}$ is a QFSH-algebra. Conversely, if B is a QFSH-algebra and m is its maximal ideal then the h -adic completion of $\sum_n h^{-n} m^n$ is a QUE-algebra.

DEFINITION. The QUE-dual of a QUE-algebra A is the QUE-algebra corresponding to A^* in the above sense (or, which is the same, the dual of the QFSH-algebra corresponding to A).

It can be shown that the classical limit of the QUE-dual of A is a Lie bialgebra dual to the classical limit of A .

8. The square of the antipode. Let A be a quantization of a Lie bialgebra \mathfrak{g} . Denote by S the antipode of A . In general $S^2 \neq \text{id}$, but $S^2 \equiv \text{id} \pmod{h}$. Here is a description of $h^{-1}(S^2 - \text{id}) \pmod{h}$ in terms of \mathfrak{g} . Denote by D the composition of the cocommutator $\mathfrak{g} \rightarrow \mathfrak{g} \otimes \mathfrak{g}$ and the commutator $\mathfrak{g} \otimes \mathfrak{g} \rightarrow \mathfrak{g}$. It is easy to show that $D: \mathfrak{g} \rightarrow \mathfrak{g}$ is a bialgebra derivation (this means that e^{tD} is a bialgebra automorphism) and $h^{-1}(S^2 - \text{id}) \pmod{h}$ is the unique derivation of $U\mathfrak{g}$ whose restriction to \mathfrak{g} is equal to $-D/2$.

If $\mathfrak{g} = \mathfrak{a}[u]$ is the bialgebra of Example 3.3 and the scalar product on \mathfrak{a} is the Killing form then $D = d/du$. For bialgebras corresponding to nondegenerate solutions of (10), (11), D becomes equal to d/du after a suitable normalization.

9. Obstruction theory. Let \mathfrak{g} be a Lie bialgebra. Denote by $\text{Der}^i \mathfrak{g}$ the $(i+2)$ th cohomology group of $\text{Ker}(C^*(\mathfrak{g} \oplus \mathfrak{g}^*) \rightarrow C^*(\mathfrak{g}) \oplus C^*(\mathfrak{g}^*))$ where $\mathfrak{g} \oplus \mathfrak{g}^*$ is equipped with a Lie algebra structure defined in §3 and $C^*(\mathfrak{a})$ is the standard complex $0 \rightarrow \mathfrak{a}^* \rightarrow \bigwedge^2 \mathfrak{a}^* \rightarrow \bigwedge^3 \mathfrak{a}^* \rightarrow \dots$. Obviously $\text{Der}^i \mathfrak{g} = 0$ for $i < 0$. $\text{Der}^0 \mathfrak{g}$ is the set of bialgebra derivations of \mathfrak{g} . It can be shown that (1) if two quantizations of \mathfrak{g} coincide modulo h^n then their difference $\pmod{h^{n+1}}$ "belongs" to $\text{Der}^1 \mathfrak{g}$, and (2) if $\text{Der}^2 \mathfrak{g} = 0$ then \mathfrak{g} has a quantization. I do not know whether every Lie bialgebra can be quantized.

10. Coboundary, triangular, and quasitriangular Hopf algebras. A pair (A, R) consisting of a Hopf algebra A and an invertible element $R \in A \otimes A$ will be called a *coboundary Hopf algebra* if

$$\Delta'(a) = R\Delta(a)R^{-1}, \quad a \in A, \quad (20)$$

and $R^{12}R^{21} = 1$, $R^{12} \cdot (\text{id} \otimes \text{id})(R) = R^{23} \cdot (\text{id} \otimes \Delta)(R)$, $(\varepsilon \otimes \varepsilon)(R) = 1$. Here $\Delta' = \sigma \circ \Delta$, $\sigma: A \otimes A \rightarrow A \otimes A$, $\sigma(x \otimes y) = y \otimes x$ and the symbols R^{12}, R^{21}, R^{23} have the following meaning: if $R = \sum_i a_i \otimes b_i$ then

$$R^{21} = \sum_i b_i \otimes a_i, \quad R^{23} = \sum_i 1 \otimes a_i \otimes b_i, \quad R^{12} = \sum_i a_i \otimes b_i \otimes 1$$

(and sometimes $R^{12} = R$). (A, R) is called a *quasitriangular Hopf algebra* if R satisfies (20) and the equations

$$(\Delta \otimes \text{id})(R) = R^{13}R^{23}, \quad (\text{id} \otimes \Delta)(R) = R^{13}R^{12}. \quad (21)$$

A quasitriangular Hopf algebra is said to be *triangular* if $R^{12}R^{21} = 1$. By definition, a coboundary (resp. triangular, quasitriangular) QUE-algebra is a

coboundary (triangular, quasitriangular) Hopf algebra (A, R) such that A is a QUE-algebra and $R \equiv 1 \pmod{h}$.

It can be shown that if (A, R) is a coboundary (triangular, quasitriangular) QUE-algebra, $r = h^{-1}(R-1) \pmod{h}$ and \mathfrak{g} is the classical limit of A then $r \in \mathfrak{g} \otimes \mathfrak{g}$ (a priori $r \in U\mathfrak{g} \otimes U\mathfrak{g}$) and (\mathfrak{g}, r) is a coboundary (triangular, quasitriangular) Lie bialgebra. There are also other arguments in favor of the above definitions.

It is easy to show that (1) a triangular Hopf algebra is a coboundary Hopf algebra, (2) if (A, R) is a quasitriangular QUE-algebra and $\bar{R} = (R^{12}R^{21})^{-1/2}R$ then (A, \bar{R}) is a coboundary QUE-algebra, (3) if (A, R) is quasitriangular then R satisfies the QYBE, i.e.

$$R^{12}R^{13}R^{23} = R^{23}R^{13}R^{12},$$

(4) if (A, R) is a coboundary QUE-algebra and R satisfies the QYBE then (A, R) is triangular, (5) equations (21) mean that the mapping $A^* \rightarrow A$ given by $l \mapsto (l \otimes \text{id})(R)$ is an algebra homomorphism and a coalgebra antihomomorphism.

To understand the various types of Hopf algebras one may use the Tannaka-Krein approach [56], i.e., instead of considering a Hopf algebra A one may consider the category Rep_A of its representations together with the forgetful functor $F: \text{Rep}_A \rightarrow \{\text{vector spaces}\}$. Recall that a representation of a Hopf algebra A is a representation of A as an algebra, and the tensor product of representations $\rho_1: A \rightarrow \text{End } V_1$ and $\rho_2: A \rightarrow \text{End } V_2$ is equal to the composition $A \xrightarrow{\Delta} A \otimes A \xrightarrow{\rho_1 \otimes \rho_2} \text{End}(V_1 \otimes V_2)$. For any A the functor $\otimes: \text{Rep}_A \times \text{Rep}_A \rightarrow \text{Rep}_A$ is associative. More precisely, for this functor there is an associativity constraint [56] compatible with F . In general the A -modules $V_1 \otimes V_2$ and $V_2 \otimes V_1$ are not isomorphic, but if (20) holds then R defines an isomorphism between them. Moreover, if $R^{12}R^{21} = 1$ we obtain a commutativity constraint [56] for $\otimes: \text{Rep}_A \times \text{Rep}_A \rightarrow \text{Rep}_A$. It is not compatible with F unless $R = 1$. The condition (21) is the compatibility of the commutativity and the associativity constraints [56]. So if (A, R) is triangular then Rep_A is a tensor category [56], but F is not a tensor functor unless $R = 1$. V. V. Lyubashenko's result [57] means that a tensor category C with a functor $F: C \rightarrow \{\text{vector spaces}\}$ having the above properties is more or less Rep_A for a unique triangular Hopf algebra A .

If (A, Δ_0) is a cocommutative Hopf algebra and ϕ is an invertible element of $A \otimes A$ such that $\phi^{12} \cdot (\Delta_0 \otimes \text{id})(\phi) = \phi^{23} \cdot (\text{id} \otimes \Delta_0)(\phi)$ then we can obtain a triangular Hopf algebra (A, Δ, R) by setting $\Delta(a) = \phi \Delta_0(a) \phi^{-1}$, $R = \phi^{21}(\phi^{12})^{-1}$. It can be shown that all triangular QUE-algebras can be obtained in this way (this is natural from the Tannaka-Krein point of view). In particular, a triangular QUE-algebra is isomorphic (as an algebra) to a universal enveloping algebra.

Finally I am going to mention a version of formula (20) which plays a crucial role in the QISM (e.g., see [71]) and which was in fact the origin of (20). Given a representation $\rho: A \rightarrow \text{Mat}(n, \mathbb{C})$, its matrix elements $t_{ij} \in A^*$ satisfy the commutation relations $R_\rho \tilde{T} \tilde{T} = \tilde{T} \tilde{T} R_\rho$, where $R_\rho = (\rho \otimes \rho)(R)$,

$T = (t_{ij})$, $\tilde{T} = T \otimes E_n$, $\tilde{\tilde{T}} = E_n \otimes T$ (E_n is the unit matrix). Moreover, for a family of representations $\rho_\lambda: A \rightarrow \text{Mat}(n, \mathbb{C})$, $\lambda \in \mathbb{C}$, we have

$$R(\lambda, \mu) \tilde{T}(\lambda) \tilde{\tilde{T}}(\mu) = \tilde{\tilde{T}}(\mu) \tilde{T}(\lambda) R(\lambda, \mu), \quad (22)$$

where $R(\lambda, \mu) = (\rho_\lambda \otimes \rho_\mu)(R)$, $T(\lambda) = (t_{ij}(\lambda))$, $t_{ij}(\lambda)$ are the matrix elements of ρ_λ . The more traditional view on (22) is to start from a function $R(\lambda, \mu) \in \text{End}(\mathbb{C}^n \otimes \mathbb{C}^n)$ and consider the algebra B with generators $t_{ij}(\lambda)$, $\lambda \in \mathbb{C}$, $1 \leq i, j \leq n$ and defining relations (22). Note that the formula

$$\Delta(t_{ij}(\lambda)) = \sum_k t_{ik}(\lambda) \otimes t_{kj}(\lambda)$$

defines a Hopf algebra structure on B (in the weak sense, i.e. without antipode), and if $R(\lambda, \mu) = (\rho_\lambda \otimes \rho_\mu)(R)$ for some $A, \{\rho_\lambda\}$ and $R \in A \otimes A$ such that (20) holds, then there is a natural homomorphism $B \rightarrow A$. So Hopf algebras have already appeared implicitly in [71]. Precise statements concerning the connection between triangular Hopf algebras and the QYBE can be found in [57, 79].

11. QISM from the Hopf algebra point of view. The QISM is a method for constructing and studying integrable quantum systems, which was created in 1978–1979 by L. D. Faddeev, E. K. Sklyanin, and L. A. Takhtajan [69–71, 37]. It is closely related to the classical inverse scattering method, Baxter’s works on exactly solvable models of classical statistical mechanics (e.g. see [58]), and one-dimensional relativistic factorized scattering theory (C. N. Yang, A. B. Zamolodchikov, ...). Here is an exposition of the basic construction of the QISM, which slightly differs from the standard expositions [70–72, 36, 37], where Hopf algebras occur implicitly and (22) is used instead of (20).

Let A be a Hopf algebra, $C = \{l \in A^* \mid l(ab) = l(ba) \text{ for any } a, b \in A\}$. Then C is a subalgebra of A^* . Now suppose that (20) holds. Then it is easy to show that C is commutative (this is the quantum analogue of (9)). Fix a representation $\pi: A^* \rightarrow \text{End } V$ (“the quantum L -operator”) and a number N (“the number of sites of a 1-dimensional lattice”). Elements of $V^{\otimes N}$ are considered as quantum states. The image of C under the representation $\pi^{\otimes N}: A^* \rightarrow \text{End } V^{\otimes N}$ is a commutative algebra of observables. Physicists usually fix an element $H \in A^*$ and consider $\pi^{\otimes N}(H)$ as a Hamiltonian.

If (A, R) is quasitriangular then R induces a homomorphism $A^* \rightarrow A$ and therefore a representation of A induces that of A^* . Usually only those representations of A^* are considered which come from representations of A . This means that instead of $\pi^{\otimes N}$ and C one works with $\rho^{\otimes N}$ and C' , where ρ is a representation of A and C' is the image of C in A , i.e. $C' = \{(l \otimes \text{id})(R) \mid l \in C\}$. To use this scheme one has to construct a quasitriangular (A, R) and as many representations of A as possible (these representations can be taken as ρ , and their characters belong to C). The Hopf algebras A used in the QISM are quantizations of Lie bialgebras G_S of Example 3.4 (i.e. of Lie bialgebras corresponding to

nondegenerate solutions of the CYBE). Yangians and quantized affine algebras are typical examples. On these algebras there exists not a quasitriangular but a "pseudotriangular" structure (see §§12,13), but this suffices to use the above scheme.

Deeper aspects of the QISM, such as the description of the spectrum of $\rho^{\otimes N}(C')$ and the limit $N \rightarrow \infty$ (see [65, 36, 71]), are not yet understood from the Hopf algebra point of view.

12. The pseudotriangular structure on Yangians and rational solutions of the QYBE. Recall that $Y(\mathfrak{a})$ is obtained by quantizing the bialgebra $\mathfrak{g} = \mathfrak{a}[u]$ of Example 3.3 and setting $\hbar = 1$. Since \mathfrak{g} is pseudotriangular it is natural to expect $Y(\mathfrak{a})$ to be pseudotriangular too. To be more precise let us consider the operators $T_\lambda: \mathfrak{g} \rightarrow \mathfrak{g}$ given by $T_\lambda(f(u)) = f(u + \lambda)$. Note that though $r = t/(u - v)$ does not belong to $\mathfrak{g} \otimes \mathfrak{g}$, the expression

$$(T_\lambda \otimes \text{id})r = t/(u + \lambda - v) = \sum_{k=0}^{\infty} t(v - u)^k \lambda^{-k-1}$$

is a power series in λ^{-1} whose coefficients belong to $\mathfrak{g} \otimes \mathfrak{g}$. Now define the automorphism $T_\lambda: Y(\mathfrak{a}) \rightarrow Y(\mathfrak{a})$ by the formulas $T_\lambda|_{\mathfrak{a}} = \text{id}$, $T_\lambda(J(a)) = J(a) + \lambda a$ for $a \in \mathfrak{a}$. It is natural to expect that there exists a formal series

$$\mathcal{R}(\lambda) = 1 + \sum_{k=1}^{\infty} \mathcal{R}_k \lambda^{-1}, \quad \mathcal{R}_k \in Y(\mathfrak{a}) \otimes Y(\mathfrak{a}),$$

which behaves as if $\mathcal{R}(\lambda)$ were equal to $(T_\lambda \otimes \text{id})R$ for some R satisfying (20), (21) and the equation $R^{12}R^{21} = 1$. This is really so [28]. More precisely, there is a unique $\mathcal{R}(\lambda)$ such that $(\Delta \otimes \text{id})\mathcal{R}(\lambda) = \mathcal{R}^{13}(\lambda)\mathcal{R}^{23}(\lambda)$ and $(T_\lambda \otimes \text{id})\Delta'(a) = \mathcal{R}(\lambda)((T_\lambda \otimes \text{id})\Delta(a))\mathcal{R}(\lambda)^{-1}$ for $a \in Y(\mathfrak{a})$. Besides, $\mathcal{R}(\lambda_1 - \lambda_2)$ satisfies the QYBE, i.e.

$$\begin{aligned} \mathcal{R}^{12}(\lambda_1 - \lambda_2)\mathcal{R}^{13}(\lambda_1 - \lambda_3)\mathcal{R}^{23}(\lambda_2 - \lambda_3) \\ = \mathcal{R}^{23}(\lambda_2 - \lambda_3)\mathcal{R}^{13}(\lambda_1 - \lambda_3)\mathcal{R}^{12}(\lambda_1 - \lambda_2). \end{aligned}$$

Moreover, $\mathcal{R}^{12}(\lambda)\mathcal{R}^{21}(-\lambda) = 1$, $(T_\mu \otimes T_\nu)\mathcal{R}(\lambda) = \mathcal{R}(\lambda + \mu - \nu)$, and the coefficient \mathcal{R}_1 is equal to t .

By the way, we are lucky that $Y(\mathfrak{a})$ is pseudotriangular and not triangular: otherwise $Y(\mathfrak{a})$ would be isomorphic (as an algebra) to a universal enveloping algebra and life would be dull.

Given a representation $\rho: Y(\mathfrak{a}) \rightarrow \text{Mat}(n, \mathbb{C})$ we obtain a matrix solution $R(\lambda_1 - \lambda_2) = (\rho \otimes \rho)(\mathcal{R}(\lambda_1 - \lambda_2))$ of the QYBE. It is of the form

$$R(\lambda_1 - \lambda_2) = 1 + (\rho \otimes \rho)(t) \cdot (\lambda_1 - \lambda_2)^{-1} + \sum_{k=2}^{\infty} A_k (\lambda_1 - \lambda_2)^{-k}. \quad (23)$$

If ρ is irreducible then $R(\lambda)$ is a rational function up to a scalar factor [28].

Now let us temporarily forget about $Y(\mathfrak{a})$ and adopt the point of view of a person who wants to construct matrix solutions of the QYBE. According to [45]

it is natural to look for solutions of the form

$$R(\lambda_1 - \lambda_2, h) = 1 + h(\rho \otimes \rho)(r(\lambda_1 - \lambda_2)) + \sum_{k=2}^{\infty} h^k a_k(\lambda_1 - \lambda_2),$$

where ρ is a representation of \mathfrak{a} and r is an $\mathfrak{a} \otimes \mathfrak{a}$ -valued solution of the CYBE. The simplest r is given by $r(\lambda) = t/\lambda$. Since this r is homogeneous it is natural to require that $R(\alpha\lambda, \alpha h) = R(\lambda, h)$. In this case $R(\lambda, h)$ is uniquely determined by $R(\lambda, 1)$, which is of the form (23). So the natural problem is to describe for a given $\rho: \mathfrak{a} \rightarrow \text{Mat}(n, \mathbb{C})$ all the solutions of the QYBE of the form (23). Theorem [28]: *if ρ is irreducible then all such solutions are of the form $(\tilde{\rho} \otimes \tilde{\rho})(\mathcal{R}(\lambda_1 - \lambda_2))$ for some representation $\tilde{\rho}: Y(\mathfrak{a}) \rightarrow \text{Mat}(n, \mathbb{C})$ such that $\tilde{\rho}|_{\mathfrak{a}} = \rho$.*

So it is clear that the problem of describing all the irreducible finite-dimensional representations of $Y(\mathfrak{a})$ is very important. It was solved for $\mathfrak{a} = \mathfrak{sl}(2)$ by V. E. Korepin and V. O. Tarasov [59–61]. For classical \mathfrak{a} specialists in the QISM have constructed by more or less elementary means a lot of R -matrices of the form (23) (see, for instance, [62, 63] and the Appendix of [45]) and therefore a lot of representations of $Y(\mathfrak{a})$. Some results concerning this problem were obtained in [28, 64]. In [54] a parametrization of the set of irreducible finite-dimensional representations of $Y(\mathfrak{a})$ (in the spirit of Cartan's highest weight theorem) is obtained using the second presentation of $Y(\mathfrak{a})$ (see §6) and the results for $Y(\mathfrak{sl}(2))$.

$Y(\mathfrak{a})$ has two “large” commutative subalgebras: the “Cartan” subalgebra generated by φ_{ik} (see §6) and the subalgebra \mathcal{A} generated by the coefficients of $(l \otimes \text{id})\mathcal{R}(\lambda)$, $l \in C$, where $C = \{l \in (Y(\mathfrak{a}))^* \mid l(ab) = l(ba) \text{ for any } a, b \in Y(\mathfrak{a})\}$. It is desirable to describe the spectra of these subalgebras in irreducible representations of $Y(\mathfrak{a})$. For \mathcal{A} this problem is especially important (see §11). For classical \mathfrak{a} and at least some representations of $Y(\mathfrak{a})$ the spectrum of \mathcal{A} is known [65–68]. The analogy between the structure of the Bethe equations and the corresponding Cartan matrices (N. Yu. Reshetikhin [68]) suggests that the spectrum of \mathcal{A} can be described for all \mathfrak{a} .

Finally, I am going to mention a realization of $Y(\mathfrak{a})$ which is often useful and which appeared much earlier than the general definition of $Y(\mathfrak{a})$. Fix a nontrivial irreducible representation $\rho: Y(\mathfrak{a}) \rightarrow \text{Mat}(n, \mathbb{C})$. Set $R(\lambda) = (\rho \otimes \rho)(\mathcal{R}(\lambda))$. Then $Y(\mathfrak{a})$ is almost isomorphic to the Hopf algebra A_ρ with generators $t_{ij}^{(k)}$, $1 \leq i, j \leq n$, $k = 1, 2, \dots$, defining relations

$$\tilde{T}(\lambda) \tilde{T}(\mu) R(\mu - \lambda) = R(\mu - \lambda) \tilde{T}(\mu) \tilde{T}(\lambda)$$

and comultiplication

$$\Delta(t_{ij}(\lambda)) = \sum_k t_{ik}(\lambda) \otimes t_{kj}(\lambda),$$

where $t_{ij}(\lambda) = \delta_{ij} + \sum_{k=1}^{\infty} t_{ij}^{(k)} \lambda^{-k}$, $T(\lambda) = (t_{ij}(\lambda))$, $\tilde{T}(\lambda) = T(\lambda) \otimes E_n$, $\tilde{\tilde{T}}(\lambda) = E_n \otimes T(\lambda)$. More precisely, there is an epimorphism $A_\rho \rightarrow Y(\mathfrak{a})$ given by $T(\lambda) \mapsto (\text{id} \otimes \rho)(\mathcal{R}(\lambda))$, and to obtain $Y(\mathfrak{a})$ from A_ρ one has to add an auxiliary relation

of the form $c(\lambda) = 1$, where $c(\lambda) \in A_\rho[[\lambda^{-1}]]$, $\Delta(c(\lambda)) = c(\lambda) \otimes c(\lambda)$, and $[a, c(\lambda)] = 0$ for any $a \in A_\rho$. For instance, if $\mathfrak{a} = \mathfrak{sl}(n)$ and $\rho: Y(\mathfrak{a}) \rightarrow \text{Mat}(n, \mathbb{C})$ is the "vector" representation then $c(\lambda)$ is the "quantum determinant" of $T(\lambda)$ in the sense of [73] and $R(\lambda)$ is proportional to the Yang R -matrix $1 + \lambda^{-1}\sigma$, where $\sigma: \mathbb{C}^n \otimes \mathbb{C}^n \rightarrow \mathbb{C}^n \otimes \mathbb{C}^n$, $\sigma(x \otimes y) = y \otimes x$.

13. The double. Trigonometric solutions of the QYBE. If \mathfrak{g} is a Lie bialgebra denote by $D(\mathfrak{g})$ the space $\mathfrak{g} \oplus \mathfrak{g}^*$ with the Lie algebra structure defined in §3. It can be shown that (1) the image r of the canonical element of $\mathfrak{g} \otimes \mathfrak{g}^*$ under the embedding $\mathfrak{g} \otimes \mathfrak{g}^* \hookrightarrow D(\mathfrak{g}) \otimes D(\mathfrak{g})$ satisfies (7) and therefore $(D(\mathfrak{g}), r)$ is a quasitriangular Lie bialgebra, and (2) the embedding $\mathfrak{g} \hookrightarrow D(\mathfrak{g})$ is a Lie bialgebra homomorphism and the embedding $\mathfrak{g}^* \hookrightarrow D(\mathfrak{g})$ is an algebra homomorphism and coalgebra antihomomorphism. $D(\mathfrak{g})$ is called the *double* of \mathfrak{g} . This construction has important applications [43].

Now let us define the *quantum double*. Let A be a Hopf algebra. Denote by A° the algebra A^* with the opposite comultiplication. It can be shown that there is a unique quasitriangular Hopf algebra $(D(A), R)$ such that (1) $D(A)$ contains A and A° as Hopf subalgebras, (2) R is the image of the canonical element of $A \otimes A^\circ$ under the embedding $A \otimes A^\circ \hookrightarrow D(A) \otimes D(A)$, and (3) the linear mapping $A \otimes A^\circ \rightarrow D(A)$ given by $a \otimes b \mapsto ab$ is bijective. As a vector space, $D(A)$ can be identified with $A \otimes A^\circ$, and the Hopf algebra structure on $A \otimes A^\circ$ can be found from the commutation relations

$$e_s e^t = \mu_s^{kjn} m_{plk}^t \sigma_n^p e^l e_j \quad \text{and} \quad e^t e_s = \mu_s^{njk} m_{klp}^t \sigma_n^p e_j e^l,$$

which follow from (1)–(3). Here $\{e_s\}$ and $\{e^t\}$ are dual bases of A and A° , while σ, m , and μ are the matrices of the skew antipode $A \rightarrow A$, the multiplication map $A \otimes A \otimes A \rightarrow A$, and the comultiplication map $A \rightarrow A \otimes A \otimes A$. If A is a QUE-algebra we will understand A° in the QUE-sense. Then $D(A)$ is a QUE-algebra, too. If \mathfrak{g} is the classical limit of A then $(D(\mathfrak{g}), r)$ is the classical limit of $(D(A), R)$.

Let $\mathfrak{g}, \mathfrak{b}_+, \mathfrak{b}_-, \mathfrak{h}$ denote the same objects as in Example 3.2. Then $D(\mathfrak{b}_+) = \mathfrak{g} \times \mathfrak{h}$ with the trivial Lie bialgebra structure on \mathfrak{h} . The quasitriangular (or pseudoquasitriangular) structure on \mathfrak{g} (see §4) is defined by the image of $r \in D(\mathfrak{b}_+) \otimes D(\mathfrak{b}_+)$ under the natural mapping $D(\mathfrak{b}_+) \otimes D(\mathfrak{b}_+) \rightarrow \mathfrak{g} \otimes \mathfrak{g}$. It can be shown that in the quantum case the situation is quite similar. Denote by $U_{\mathfrak{h}} \mathfrak{b}_+$ the subalgebra of $U_{\mathfrak{h}} \mathfrak{g}$ generated by \mathfrak{h} and X_i^+ . Then $U_{\mathfrak{h}} \mathfrak{b}_- = (U_{\mathfrak{h}} \mathfrak{b}_+)^{\circ}$ and $D(U_{\mathfrak{h}} \mathfrak{b}_+) = U_{\mathfrak{h}} \mathfrak{g} \otimes U_{\mathfrak{h}}$. Denote by \bar{R} the image of $R \in D(U_{\mathfrak{h}} \mathfrak{b}_+) \otimes D(U_{\mathfrak{h}} \mathfrak{b}_+)$ under the natural mapping $D(U_{\mathfrak{h}} \mathfrak{b}_+) \otimes D(U_{\mathfrak{h}} \mathfrak{b}_+) \rightarrow U_{\mathfrak{h}} \mathfrak{g} \otimes U_{\mathfrak{h}} \mathfrak{g}$. Then \bar{R} defines a quasitriangular structure on $U_{\mathfrak{h}} \mathfrak{g}$ if $\dim \mathfrak{g} < \infty$ and a "pseudoquasitriangular" structure if $\dim \mathfrak{g} = \infty$.

If $\mathfrak{g} = \mathfrak{sl}(2)$ with the scalar product $\text{Tr}(XY)$ then

$$\bar{R} = \sum_{k=0}^{\infty} h^k Q_k(h) \left\{ \exp \frac{h}{4} [H \otimes H + k(H \otimes 1 - 1 \otimes H)] \right\} (X^+)^k \otimes (X^-)^k,$$

where $Q_k(h) = e^{-kh/2} \prod_{r=1}^k (e^h - 1)/(e^{rh} - 1)$. In general,

$$\bar{R} = \sum_{\beta \in \mathbf{Z}_+^r} \left\{ \exp h \left[\frac{1}{2} t_0 + \frac{1}{4} (H_\beta \otimes 1 - 1 \otimes H_\beta) \right] \right\} P_\beta,$$

where t_0 is the element of $\mathfrak{h} \otimes \mathfrak{h}$ corresponding to the canonical scalar product on \mathfrak{h} , $\mathbf{Z}_+ = \{n \in \mathbf{Z} \mid n \geq 0\}$, r is the rank of \mathfrak{g} , for every $\beta = (\beta_1, \beta_2, \dots) \in \mathbf{Z}^r$ the symbol H_β denotes $\sum_i \beta_i H_i$, and $P_\beta \in U_h \mathfrak{b}_+ \otimes U_h \mathfrak{b}_-$ is a polynomial in $u_i = X_i^+ \otimes 1$ and $v_i = 1 \otimes X_i^-$ homogeneous with respect to each variable and such that $\deg_{u_i} P_\beta = \deg_{v_i} P_\beta = \beta_i$. In addition: (1) $P_0 = 1$, (2) $P_\beta \equiv 0 \pmod{h^{k_\beta}}$, $P_\beta \not\equiv 0 \pmod{h^{k_\beta-1}}$, where $k_\beta = \inf\{n \mid H_\beta \text{ is representable as a sum of } n \text{ positive roots of } \mathfrak{g}\}$, and (3) the coefficients of $h^{-|\beta|} P_\beta$ are rational functions of e^{ch} for $c \in \mathbf{C}$, where $|\beta| = \sum_i \beta_i$.

If $\dim \mathfrak{g} < \infty$ then $R_\rho \stackrel{\text{def}}{=} (\rho \otimes \rho)(\bar{R})$ makes sense and satisfies the QYBE for any representation $\rho: U_h \mathfrak{g} \rightarrow \text{Mat}(n, \mathbf{C}[[h]])$. For instance, if $\mathfrak{g} = \mathfrak{sl}(n)$ and ρ is the representation mentioned in §7 then

$$R_\rho = e^{-h/2n} \left\{ \sum_{i \neq j} e_{ii} \otimes e_{jj} + e^{h/2} \sum_i e_{ii} \otimes e_{ii} + (e^{h/2} - e^{-h/2}) \sum_{i < j} e_{ij} \otimes e_{ji} \right\},$$

where e_{ij} are the matrix units (this formula is implicit in [52]). Note that the relations (16)–(19) simply mean that $R_\rho \tilde{T} \tilde{T} = \tilde{T} \tilde{T} R_\rho$, where $T = (\rho_{ij})$.

Now let \mathfrak{g} be affine and set $\mathfrak{g}' = [\mathfrak{g}, \mathfrak{g}]/(\text{the center of } \mathfrak{g})$. Denote by R' the analogue of \bar{R} for $U_h \mathfrak{g}'$. Suppose that a representation $\rho: U_h \mathfrak{g}' \rightarrow \text{Mat}(n, \mathbf{C}[[h]])$ is given. Strictly speaking, the expression $(\rho \otimes \rho)(R')$ is meaningless. But if we define $T_\lambda \in \text{Aut } U_h \mathfrak{g}$ setting $T_\lambda(X_i^\pm) = \lambda^{\pm 1} X_i^\pm$ and put $R'(\lambda) = (T_\lambda \otimes \text{id})R'$ then $R_\rho(\lambda) \stackrel{\text{def}}{=} (\rho \otimes \rho)(R'(\lambda))$ is a well-defined formal series in λ such that

$$R_\rho^{12}(\lambda_1/\lambda_2) R_\rho^{13}(\lambda_1/\lambda_3) R_\rho^{23}(\lambda_2/\lambda_3) = R_\rho^{23}(\lambda_2/\lambda_3) R_\rho^{13}(\lambda_1/\lambda_3) R_\rho^{12}(\lambda_1/\lambda_2).$$

Using an idea of M. Jimbo [50] it can be proved that if ρ is irreducible then $R_\rho(\lambda)$ is a rational function of λ up to a scalar factor. Moreover, for a given ρ this rational function can be found by solving certain linear equations [50]. Substituting $\lambda = e^u$ we obtain trigonometric solutions of the QYBE. Explicit formulas for the solutions corresponding to the classical \mathfrak{g} and some ρ were found by elementary methods [48, 74–77].

14. Final remarks. Because of the lack of space I cannot discuss here the relation between $Y(\mathfrak{sl}(n))$, $U_h \mathfrak{sl}(n)$ and Hecke algebras (see [52, 64]), between Yangians and Hermitian symmetric spaces (see [62] and [28, Theorem 7]), between $Y(\mathfrak{sl}(n))$ and representations of infinite-dimensional classical groups (see [80]). By the same reason I cannot discuss compact quantum groups [81–83] as well as the attempts to generalize the notion of superalgebra, which are based on the QYBE [78] or triangular Hopf algebras [79].

REFERENCES

1. J. W. Milnor and J. C. Moore, *On the structure of Hopf algebras*, Ann. of Math. (2) **81** (1965), 211–264.
2. M. E. Sweedler, *Hopf algebras*, Mathematical Lecture Note Series, Benjamin, N.Y., 1969.
3. E. Abe, *Hopf algebras*, Cambridge Tracts in Math., No. 74, Cambridge Univ. Press, Cambridge-New York, 1980.
4. G. M. Bergman, *Everybody knows what a Hopf algebra is*, Contemp. Math. **43** (1985), 25–48.
5. M. E. Sweedler, *Integrals for Hopf algebras*, Ann. of Math. (2) **89** (1969), 323–335.
6. R. G. Larson, *Characters of Hopf algebras*, J. Algebra **17** (1971), 352–368.
7. W. C. Waterhouse, *Antipodes and group-likes in finite Hopf algebras*, J. Algebra **37** (1975), 290–295.
8. D. E. Radford, *The order of the antipode of a finite dimensional Hopf algebra is finite*, Amer. J. Math. **98** (1976), 333–355.
9. H. Shigano, *A correspondence between observable Hopf ideals and left coideal subalgebras*, Tsukuba J. Math. **1** (1977), 149–156.
10. E. J. Taft and R. L. Wilson, *There exist finite dimensional Hopf algebras with antipodes of arbitrary even order*, J. Algebra **62** (1980), 283–291.
11. H. Shigano, *On observable and strongly observable Hopf ideals*, Tsukuba J. Math. **6** (1982), 127–150.
12. B. Pareigis, *A noncommutative noncocommutative Hopf algebra in "nature"*, J. Algebra **70** (1981), 356–374.
13. G. I. Kac, *Generalizations of the group duality principle*, Dokl. Akad. Nauk SSSR **138** (1961), 275–278. (Russian)
14. —, *Compact and discrete ring groups*, Ukrainian Math. J. **14** (1962), 260–270. (Russian)
15. —, *Ring groups and the duality principle*, Proc. Moscow Math. Soc. **12** (1963), 259–303; **13** (1965), 84–113. (Russian)
16. G. I. Kac and V. G. Palyutkin, *An example of a ring group generated by Lie groups*, Ukrainian Math. J. **16** (1964), 99–105. (Russian)
17. —, *Finite ring groups*, Proc. Moscow Math. Soc. **15** (1966), 224–261. (Russian)
18. G. I. Kac, *Group extensions which are ring groups*, Mat. Sb. **76** (1968), 473–496. (Russian)
19. —, *Some arithmetical properties of ring groups*, Funktsional Anal. i Prilozhen. **6** (1972), no. 2, 88–90. (Russian)
20. G. I. Kac and L. I. Vainerman, *Non-unimodular ring groups and Hopf-von Neumann algebras*, Mat. Sb. **94** (1974), 194–225. (Russian)
21. M. Enock and J.-M. Schwartz, *Une dualité dans les algèbres de von Neumann*, Bull. Soc. Math. France, mém. No 44, 1975.
22. J.-M. Schwartz, *Relations entre "ring groups" et algèbres de Kac*, Bull. Sci. Math. (2) **100** (1976), 289–300.
23. M. Enock, *Produit croisé d'une algèbre de von Neumann par une algèbre de Kac*, J. Funct. Anal. **26** (1977), 16–47.
24. M. Enock and J.-M. Schwartz, *Produit croisé d'une algèbre de von Neumann par une algèbre de Kac. II*, Publ. Res. Inst. Math. Sci. **16** (1980), 189–232.
25. J. De Cannière, M. Enock, and J.-M. Schwartz, *Algèbres de Fourier associées à une algèbre de Kac*, Math. Ann. **245** (1979), 1–22.
26. —, *Sur deux résultats d'analyse harmonique non-commutative: une application de la théorie des algèbres de Kac*, J. Operator Theory **5** (1981), 171–194.
27. V. G. Drinfeld, *Hamiltonian structures on Lie groups, Lie bialgebras and the geometric meaning of the classical Yang-Baxter equations*, Dokl. Akad. Nauk SSSR **268** (1983), 285–287. (Russian)
28. —, *Hopf algebras and the quantum Yang-Baxter equation*, Dokl. Akad. Nauk SSSR **283** (1985), 1060–1064. (Russian)

29. I. M. Gelfand and I. V. Cherednik, *Abstract Hamiltonian formalism for classical Yang-Baxter bundles*, Uspekhi Mat. Nauk **38** (1983), no. 3, 3–21. (Russian)
30. I. V. Cherednik, *Bäcklund-Darboux transformations for classical Yang-Baxter bundles*, Funktsional. Anal. i Prilozhen. **17** (1983), no. 2, 88–89. (Russian)
31. —, *On the definition of τ -functions for generalized affine Lie algebras*, Funktsional. Anal. i Prilozhen. **17** (1983), no. 3, 93–95. (Russian)
32. I. M. Gelfand and I. Ya. Dorfman, *Hamiltonian operators and the classical Yang-Baxter equation*, Funktsional. Anal. i Prilozhen. **16** (1982), no. 4, 1–9. (Russian)
33. A. Weinstein, *The local structure of Poisson manifolds*, J. Diff. Geometry **18** (1983), 523–557.
34. J. A. Schouten, *Über differentialkomitanten zweier kontravarianter Größen*, Nederl. Akad. Wetensch. Proc. Ser. A **43** (1940), 449–452.
35. L. D. Faddeev and L. A. Takhtajan, *Hamiltonian approach to solitons theory*, Springer-Verlag, Berlin-New York, 1987.
36. L. D. Faddeev, *Integrable models in (1+1)-dimensional quantum field theory* (Lectures in Les Houches, 1982), Elsevier Science Publishers B. V., 1984.
37. E. K. Sklyanin, *The quantum version of the inverse scattering method*, Zap. Nauchn. Sem. LOMI **95** (1980), 55–128.
38. L. D. Faddeev and N. Yu. Reshetikhin, *Hamiltonian structures for integrable models of field theory*, Theoret. Math. Phys. **56** (1983), 323–343. (Russian)
39. M. A. Semenov-Tian-Shansky, *What is the classical r -matrix*, Funktsional. Anal. i Prilozhen. **17** (1983), no. 4, 17–33. (Russian)
40. A. A. Belavin and V. G. Drinfeld, *On the solutions of the classical Yang-Baxter equations for simple Lie algebras*, Funktsional. Anal. i Prilozhen. **16** (1982), no. 3, 1–29. (Russian)
41. —, *Triangle equations and simple Lie algebras*, Soviet Scientific Reviews Section C **4** (1984), 93–165. Harwood Academic Publishers, Chur (Switzerland)-New York.
42. —, *On the classical Yang-Baxter equation for simple Lie algebras*, Funktsional. Anal. i Prilozhen. **17** (1983), no. 3, 69–70. (Russian)
43. M. A. Semenov-Tian-Shansky, *Dressing transformations and Poisson group actions*, Publ. Res. Inst. Math. Sci. **21** (1986).
44. E. K. Sklyanin, *On complete integrability of the Landau-Lifshitz equation*, LOMI preprint E-3-1979, Leningrad, 1979.
45. P. P. Kulish and E. K. Sklyanin, *On the solutions of the Yang-Baxter equations*, Zap. Nauchn. Sem. LOMI **95** (1980), 129–160. (Russian)
46. A. A. Belavin, *Discrete groups and integrability of quantum systems*, Funktsional. Anal. i Prilozhen. **14** (1980), 18–26. (Russian)
47. G. E. Andrews, *The theory of partitions*, Addison-Wesley, London, 1976.
48. P. P. Kulish and N. Yu. Reshetikhin, *The quantum linear problem for the sine-Gordon equation and higher representations*, Zap. Nauchn. Sem. LOMI **101** (1981), 101–110. (Russian)
49. E. K. Sklyanin, *On an algebra generated by quadratic relations*, Uspekhi Mat. Nauk **40** (1985), no. 2, 214.
50. M. Jimbo, *Quantum R -matrix for the generalized Toda system*, Commun. Math. Phys. **102** (1986), 537–547.
51. —, *A q -difference analogue of $U(\mathfrak{g})$ and the Yang-Baxter equation*, Lett. Math. Phys. **10** (1985), 63–69.
52. —, *A q -analogue of $U(\mathfrak{gl}(N+1))$, Hecke algebra and the Yang-Baxter equation*, Lett. Math. Phys. **11** (1986), 247–252.
53. C. N. Yang, *Some exact results for the many-body problem in one dimension with repulsive delta-function interaction*, Phys. Rev. Letters **19** (1967), 1312–1314.
54. V. G. Drinfeld, *A new realization of Yangians and quantized affine algebras*, FTINT Preprint No. 30–86, Kharkov, 1986 (to appear in Dokl. Akad. Nauk SSSR). (Russian)
55. V. G. Kac, *Infinite dimensional Lie algebras*, Birkhäuser, Boston-Basel-Stuttgart, 1983.
56. P. Deligne and J. Milne, *Tannakian categories*, Lecture Notes in Math., vol. 900, Springer-Verlag, Berlin-Heidelberg-New York, 1982, pp. 101–228.

57. V. V. Lyubashenko, *Hopf algebras and vectorsymmetries*, Uspekhi Mat. Nauk **41** (1986), no. 5, 185–186.
58. R. J. Baxter, *Partition function of the eight-vertex lattice model*, Ann. Phys. (N.Y.) **70** (1972), 193–228.
59. V. E. Korepin, *The analysis of the bilinear relation of the six-vertex model*, Dokl. Akad. Nauk SSSR **265** (1982), 1361–1364. (Russian)
60. V. O. Tarasov, *On the structure of quantum L -operators for the R -matrix of the XXZ -model*, Theoret. Math. Phys. **61** (1984), 163–173. (Russian)
61. —, *Irreducible monodromy matrices for the R -matrix of the XXZ -model and lattice local quantum Hamiltonians*, Theoret. Math. Phys. **63** (1985), 175–196. (Russian)
62. N. Yu. Reshetikhin, *Integrable models of quantum 1-dimensional magnetics with $O(n)$ and $Sp(2k)$ -symmetry*, Theoret. Math. Phys. **63** (1985), 347–366. (Russian)
63. P. P. Kulish, N. Yu. Reshetikhin, and E. K. Sklyanin, *Yang-Baxter equation and representation theory. I*, Lett. Math. Phys. **5** (1985), 393–403.
64. V. G. Drinfeld, *Degenerate affine Hecke algebras and Yangians*, Funktsional. Anal. i Prilozhen. **20** (1986), no. 1, 69–70. (Russian)
65. L. A. Takhtajan, *Quantum inverse scattering method and the algebrized matrix Bethe ansatz*, Zap. Nauchn. Sem. LOMI **101** (1981), 158–183. (Russian)
66. P. P. Kulish and N. Yu. Reshetikhin, *The generalized Heisenberg ferromagnetic and the Gross-Neveu Model*, J. Experim. Theoret. Phys. **80** (1981), 214–227. (Russian)
67. N. Yu. Reshetikhin, *Functional equation method in the theory of exactly solvable quantum systems*, J. Experim. Theoret. Phys. **84** (1983), 1190–1201. (Russian)
68. —, *Exactly solvable quantum systems on the lattice, which are connected with classical Lie algebras*, Zap. Nauchn. Sem. LOMI **123** (1983), 112–125. (Russian)
69. L. D. Faddeev and E. K. Sklyanin, *The quantum-mechanical approach to completely integrable models of field theory*, Dokl. Akad. Nauk SSSR **243** (1978), 1430–1433. (Russian)
70. L. D. Faddeev, E. K. Sklyanin, and L. A. Takhtajan, *The quantum inverse problem I*, Theoret. Math. Phys. **40** (1979), 194–220. (Russian)
71. L. D. Faddeev and L. A. Takhtajan, *The quantum inverse problem method and the XYZ Heisenberg model*, Uspekhi Mat. Nauk **34** (1979), no. 5, 13–63. (Russian)
72. L. D. Faddeev, *Quantum completely integrable models in field theory*, Soviet Scientific Reviews Section C **1** (1980), 107–155; Harwood Academic Publishers, Chur (Switzerland)-New York.
73. P. P. Kulish and E. K. Sklyanin, *Quantum spectral transform method. Recent developments*, Lecture Notes in Phys., vol. 151, Springer-Verlag, Berlin-Heidelberg-New York, 1982, pp. 61–119.
74. I. V. Cherednik, *On the method of constructing factorized S -matrices in elementary functions*, Theoret. Math. Phys. **43** (1980), 117–119. (Russian)
75. A. V. Zamolodchikov and V. A. Fateev, *A model factorized S -matrix and the integrable Heisenberg lattice with spin 1*, Yadernaya Fiz. **32** (1980), 581–590. (Russian)
76. A. G. Izergin and V. E. Korepin, *The inverse scattering approach to the quantum Shabat-Mikhailov model*, Comm. Math. Phys. **79** (1981), 303–316.
77. V. V. Bazhanov, *Trigonometric solutions of triangle equations and classical Lie algebras*, Phys. Lett. **159B** (1985), 321–324.
78. D. I. Gurevich, *Quantum Yang-Baxter equation and generalization of the formal Lie theory*, Seminar on Supermanifolds (D. A. Leites, ed.), D. Reidel (to appear).
79. V. V. Lyubashenko, *Vectorsymmetries*, ibid.
80. G. I. Olshansky, *An extension of $U(\mathfrak{g})$ for infinite dimensional Lie algebras \mathfrak{g} and Yangians*, Dokl. Akad. Nauk SSSR (to appear).
81. S. L. Woronowicz, *Twisted $SU(2)$ group. An example of a non-commutative differential calculus*, Publications RIMS Kyoto University (to appear).
82. —, *Compact matrix pseudogroups*, Preprint No. 1/87, Warsaw University, 1987.
83. L. L. Vaksman and Ya. S. Soibelman, *The algebra of functions on quantized $SU(2)$* , Funktsional. Anal. i Prilozhen. (to appear).

Beyond Affine Lie Algebras

I. B. FRENKEL

1. Introduction. The theory of groups and their representations has been recognized for a long time as a source of canonical structures for practically all branches of mathematics and theoretical physics. The theory is remarkable in its internal beauty, and every new connection strikes us with the profundity of the symmetry principle underlying so much of our knowledge. Undoubtedly the first such source of fundamental symmetries is the theory of finite-dimensional simple Lie groups, one of the most refined and harmonious chapters in mathematics. The classification of these groups and their infinitesimal counterparts, finite-dimensional simple Lie algebras, point to the primordial structures common to many fundamental objects in mathematics, namely, ADE diagrams and their symmetries. The simplest example of ADE classification, regular solids, was the pride of Greek mathematics. At the other end of the long list of ADE classifications is the class of affine Lie algebras, the simplest nontrivial subclass of infinite-dimensional Lie algebras from several different points of view (§§2–6). This fact and in particular the underlying ADE structure indicate the canonical nature of affine Lie algebras and the corresponding groups. Affine and finite-dimensional simple Lie algebras of ADE type are in one-to-one correspondence, but the former Lie algebras have more symmetrical diagrams than the latter, yielding a larger classification of the others. Many results for finite-dimensional simple Lie algebras can be best understood by the study of the corresponding affine Lie algebras. More importantly, the affine Lie algebras and the corresponding groups can also be recognized as the initial source of numerous old and new mathematical and physical theories, which appeared earlier as completely unrelated. Thus, they provide us with a unified perspective for many isolated results.

I will begin with the definition of affine Lie algebras as simple \mathbf{Z} -graded Lie algebras (§2). Then I will consider their various manifestations: as Kac-Moody algebras (§3), as loop algebras (§4), as operator algebras (§5), and as vertex algebras (§6). Different realizations lead to different analogues and consequent results as well as to different applications and generalizations. In §7 I will review some results about the Virasoro algebra, which is inseparable from the theory

of affine Lie algebras. Two of the four approaches to affine Lie algebras (§§4 and 6) lead naturally to significant generalizations which are the subjects of §§8 and 9, respectively. In nearly every section, starting from affine Lie algebras I also want to show the structures and connections *beyond affine Lie algebras*. To give a complete and at the same time, developing picture, I have included both established results and open problems. Since the relevant material recently has overgrown all possible limits I apologize to the authors whose results were omitted because of my inadvertence or lack of time.

2. Affine Lie algebras and Virasoro algebra: Characterization.

Affine Lie algebras admit a concise characterization similar to that of simple finite-dimensional Lie algebras. It was recently given in [15], and it simplifies the original characterization contained in [Kac (c)]. We will use this result as a definition of affine Lie algebras.

Let $\mathfrak{G} = \bigoplus_{n \in \mathbb{Z}} \mathfrak{G}_n$ be a complex simple \mathbb{Z} -graded Lie algebra infinite-dimensional in both directions with uniformly bounded $\dim \mathfrak{G}_n$, $n \in \mathbb{Z}$. Then \mathfrak{G} is either an affine Lie algebra or the Witt algebra.

The lifting of the condition that \mathfrak{G} be infinite-dimensional in both directions would only add finite-dimensional simple Lie algebras and a subalgebra of the Witt algebra [15]. O. Mathieu's result states that the affine Lie algebra \mathfrak{G} defined as above admits a realization as an appropriate twisting of $\mathfrak{g} \otimes \mathbb{C}[t, t^{-1}]$ for some simple finite-dimensional Lie algebra \mathfrak{g} . This realization is discussed in §4; here we will only use two general implications of this fact. First is the existence of a bilinear invariant symmetric nondegenerate form $\langle \cdot, \cdot \rangle$ on the affine Lie algebra \mathfrak{G} such that $\langle \mathfrak{G}_m, \mathfrak{G}_n \rangle = 0$ for $m + n \neq 0$ [Kac (c)], [Moody (c)]. Second is that $H^2(\mathfrak{G}) \cong \mathbb{C}$ [Garland (c)], [R. Wilson]. Let d be the derivation of \mathfrak{G} such that $d \cdot x = nx$, $x \in \mathfrak{G}_n$; then $\alpha(x, y) = n\langle x, y \rangle$, $x \in \mathfrak{G}_n$, $y \in \mathfrak{G}_{-n}$, represents a nontrivial element in $H^2(\mathfrak{G})$. We define $\hat{\mathfrak{G}}$ as a nontrivial one-dimensional central extension of \mathfrak{G} , and we consider the semidirect products $\check{\mathfrak{G}} = Cd + \mathfrak{G}$ and $\tilde{\mathfrak{G}} = Cd + \hat{\mathfrak{G}}$. One has a commutative diagram:

$$\begin{array}{ccc}
 & \hat{\mathfrak{G}} & \\
 \swarrow & & \searrow \\
 \tilde{\mathfrak{G}} & & \mathfrak{G} \\
 \searrow & & \swarrow \\
 & \check{\mathfrak{G}} &
 \end{array} \tag{1}$$

The definition of $\hat{\mathfrak{G}}, \check{\mathfrak{G}}, \tilde{\mathfrak{G}}$ does not depend on the grading of \mathfrak{G} , and we call all four Lie algebras affine Lie algebras.

The appearance of the Witt algebra in the above characterization indicates its close relation to affine Lie algebras which we will discuss in §7. We denote the Witt algebra by \mathfrak{L} . Contrary to affine Lie algebras \mathfrak{L} does not admit an invariant symmetric bilinear form; however, one has $H^2(\mathfrak{L}) \cong \mathbb{C}$ [10]. The central extension $\hat{\mathfrak{L}}$ of \mathfrak{L} is called the Virasoro algebra. $\hat{\mathfrak{L}}$ is spanned by L_m ,

$m \in \mathbf{Z}$, and the central element c with the Lie bracket

$$[L_m, L_n] = (m - n)L_{m+n} + \frac{m^3 - m}{12}\delta_{m+n,0}c, \quad m, n \in \mathbf{Z}. \quad (2)$$

3. Affine Lie algebras as Kac-Moody algebras. Affine Lie algebras have a structure theory completely analogous to that of simple finite-dimensional Lie algebras. This allows us to generalize a great number of classical results to the affine case. This approach is best formulated in terms of the more general Kac-Moody Lie algebras.

3.1. Construction via generators and relations. A finite-dimensional simple Lie algebra \mathfrak{g} of rank r can be reconstructed from the corresponding $r \times r$ Cartan matrix A by means of $3r$ generators $e_i, f_i, h_i, i = 1, \dots, r$, and standard relations [Serre]. It was discovered by [Kac (c)], [Moody (b)] and in the final form in [Gabber and Kac] that an affine Lie algebra $\hat{\mathfrak{g}}$ can be obtained in a similar fashion from a unique *affine Cartan matrix* $A = (A_{ij}), i, j \in I$, characterized by the properties (a) $A_{ii} = 2, i \in I$, (b) $-A_{ij} \in \mathbf{Z}_+$ and $A_{ij} = 0 \Leftrightarrow A_{ji} = 0, i \neq j \in I$, (c) if $A_{ij} = 0$ for $i \in I_0 \subset I, j \notin I_0$ then $I_0 = \emptyset$ or I , (d) all principal minors of A are ≥ 0 and $\det A = 0$. If we change the property (b) to (b') $A_{ij} = A_{ji} = 0, -1$ or $-2, i \neq j \in I$, then the corresponding affine Lie algebras are called simply laced by analogy with the finite-dimensional case.

The classification of affine Cartan matrices, which yields the classification of affine Lie algebras, was given in [Kac (c)], [Moody (c)], and [Macdonald (a)]. It is especially transparent in the simply laced case, when there is a one-to-one correspondence between the Cartan matrix of type Γ and its affine counterpart $\hat{\Gamma}$. The complete list in this case consists of the omnipresent ADE types mentioned in the Introduction: $A_r, r \geq 1, D_r, r \geq 4, E_r, r = 6, 7, 8$. The remaining simple and affine Lie algebras come from the twisting, determined from the symmetries of Cartan matrices, of the above types.

Cartan matrices satisfying only the conditions (a) and (b) generate the class of Kac-Moody algebras $\mathfrak{g}(A)$ [Kac (c)], [Moody (b)]. A great number of results about simple and affine Lie algebras can be obtained in this generality. In particular, V. Kac and R. Moody independently defined most of the following structures of $\mathfrak{g}(A)$: the Weyl group W , the root system Δ and its subsystem of real roots Δ_R , the bilinear symmetric invariant form $\langle \cdot, \cdot \rangle$, the canonical involution σ of $\mathfrak{g}(A)$, the compact form of $\mathfrak{g}(A)$, the "nilpotent" subalgebras $\mathfrak{n}^+, \mathfrak{n}^-$, etc.

3.2. Results on Kac-Moody algebras. Most of the results concerning Kac-Moody algebras known by 1978 were reviewed in [Kac '78] and [Lepowsky '78]. Among more recent results I would like to mention the following: (1) on structure of Kac-Moody algebras and dominant highest weight modules [Kac-Peterson (c), (e)]; (2) on construction of groups associated with Kac-Moody algebras [Tits (c)], [Mathieu (b)], [Kac-Peterson (d)]; (3) on cohomology of Kac-Moody algebras, groups, and their flag varieties [Kumar (c)], [Gutkin-Slodowy], [Kac-Peterson (j)]; (4) on singularities associated to Kac-Moody algebras and groups

[Slodowy (b), (c)]. All the results are based on the corresponding facts for finite-dimensional simple Lie algebras and groups but formulated without use of finite-dimensionality and are valid in particular for affine Lie algebras and groups. In the next section we will see how these results acquire a new meaning in the concrete realizations of affine Lie algebras. Unfortunately, very little is known about the *realization of nonaffine Kac-Moody algebras and their root multiplicities*. Some partial results were obtained in [Feingold-Frenkel] (see also §8.1).

4. Affine Lie algebras as loop algebras. In the concrete geometric realization many formal results about affine Lie algebras and groups which are analogues to the finite-dimensional case acquire the aspects of apparently different mathematical and physical theories. Among those are elliptic and modular functions, Floquet theory, Birkhoff decomposition, topology of loop spaces, Gaussian and Wiener measure, two-dimensional quantum field theory. One can reinterpret old results, find new ones, and formulate some natural problems.

4.1. *Geometric realization.* As we mentioned in §2 an affine Lie algebra admits a realization as an appropriate twisting of $\bar{\mathfrak{g}} = \mathfrak{g} \otimes \mathbb{C}[t, t^{-1}]$ for some simple finite-dimensional Lie algebra \mathfrak{g} . For simplicity we will consider only the untwisted case, which, in particular, includes the simply laced affine Lie algebras. The affine Lie algebra $\bar{\mathfrak{g}}$ is naturally isomorphic to the loop algebra of \mathfrak{g} , i.e., the Lie algebra of polynomial maps of S^1 into \mathfrak{g} . The compact form of $\bar{\mathfrak{g}}$ is identified with the maps of S^1 into the compact form of \mathfrak{g} . From now on \mathfrak{g} will denote either the compact simple Lie algebra \mathfrak{g}_c or the complex algebra $\mathfrak{g}_{\mathbb{C}}$ unless it is specified. Now the two-cocycle on $\bar{\mathfrak{g}}$ has the form

$$\alpha(x(\cdot), y(\cdot)) = \frac{\delta^2}{2\pi} \int_{S^1} \langle x(s), y'(s) \rangle ds,$$

where δ is the length of the maximal root of \mathfrak{g} , $\langle \cdot, \cdot \rangle$ denotes the Killing form on \mathfrak{g} , and $d = (1/i)d/ds$ is the derivation of $\bar{\mathfrak{g}}$. The form of the cocycle α entails the “maximal” completion of $\hat{\bar{\mathfrak{g}}}$ in the $H^{1/2}$ topology, which will be considered in §5. The adjoining of the derivation d involves naturally the completion in the C^∞ topology, which we will use in this section: $\bar{\mathfrak{g}} = C^\infty(S^1, \mathfrak{g})$.

The root decomposition is best formulated for the affine Lie algebra $\tilde{\bar{\mathfrak{g}}} = \tilde{\mathfrak{h}} + \bigoplus_{\alpha \in \tilde{\Delta}} \tilde{\bar{\mathfrak{g}}}^\alpha$. The affine Weyl group \tilde{W} and the affine root system $\tilde{\Delta}$ are realized in terms of the corresponding structures of \mathfrak{g} . In fact, one has $\tilde{W} \cong W \ltimes Q^\vee$, $\tilde{\Delta} \cup 0 \cong (\Delta \cup 0) \times \mathbb{Z}$, where Q^\vee is the dual root lattice of \mathfrak{g} [Kac (c)], [Moody (c)], [Macdonald (a)]. Note that if the rank of $\mathfrak{h}_{\mathbb{R}}$ is equal to r , then its affine counterpart $\tilde{\mathfrak{h}}_{\mathbb{R}}$ has signature $(r+1, 1)$.

Let G be a finite-dimensional simple Lie group with the Lie algebra \mathfrak{g} . Then the adjoint group \bar{G} associated with $\bar{\mathfrak{g}}$ is identified with $C^\infty(S^1, G)$. The adjoint action of \bar{G} on $\bar{\mathfrak{g}}$ differs from the pointwise adjoint action by the 1-cocycle

$$\beta(g(\cdot), x(\cdot)) = \frac{\delta^2}{4\pi} \int_{S^1} \langle x(s), g(s)^{-1} g'(s) \rangle ds.$$

The coadjoint action of \hat{G} on $\check{\mathfrak{g}} \subset \hat{\mathfrak{g}}^*$ preserves the hyperplanes $bd + \bar{\mathfrak{g}}$, $b \in \mathbf{C}$ for $\mathfrak{g} = \mathfrak{g}_{\mathbf{C}}$ and $b \in i\mathbf{R}$ for $\mathfrak{g} = \mathfrak{g}_{\mathbf{c}}$, and has the form of a gauge action $\text{Ad}^*g \cdot (bd + x) = bd + g^{-1}xg + bg^{-1}g'$. Classification of orbits of the coadjoint action follows from the Floquet theory of differential equations with periodic coefficients. The orbits are parametrized by the fundamental domain of \hat{W} , in agreement with the corresponding fact in the finite-dimensional case. It also follows that the orbit $\hat{\mathcal{O}}_x$, $x \in \check{\mathfrak{g}}$, of the coadjoint representation is naturally identified with a path space $\text{Path}(\mathcal{O}(x))$ of all C^∞ -paths in G with the initial point $e \in G$ and the final point in the conjugacy class $\mathcal{O}(x) \subset G$ depending on x [Frenkel (g)], [Segal].

By analogy with the finite-dimensional case, $\bar{\mathfrak{g}}_{\mathbf{C}}$ admits a parabolic decomposition $\mathfrak{g}^+ + \mathfrak{g}^-$, where \mathfrak{g}^+ (resp. \mathfrak{g}^-) is identified with the loops which can be analytically (resp. antianalytically) continued to the unit disc D . Coadjoint orbits of G^+ produce various classical completely integrable systems [van Moerbeke '83]. The corresponding decomposition of $\bar{G}_{\mathbf{C}}$ can be identified as the Birkhoff decomposition $G^+(\hat{W}/W)G^-$ [Garland and Raghunathan], [Pressley (a)] and is related to the Riemann-Hilbert problem.

The explicit construction of the affine Lie group \hat{G} is not as straightforward as in the case of its Lie algebra $\hat{\mathfrak{g}}$, since the central extension is topologically nontrivial. Nevertheless one can in principle obtain an explicit construction of \hat{G} as a factor group, which was recently completed in [16b] and has a long history in the physical literature (see, e.g., [20a]). We notice first that $\bar{G} = G^D/\mathcal{G}$, where $G^D = C^\infty(D, G)$, $\mathcal{G} = \{g \in G^D : g|_{S^1} = \text{Id}\} \cong C_{\text{based}}^\infty(S^2, G)$. We will call \mathcal{G} a gauge group. Let \hat{G}^D be the central extension of G^D defined by the cocycle $e^{il\gamma(g_1, g_2)}$, where

$$\gamma(g_1, g_2) = \frac{1}{4\pi\delta^2} \int_D \text{tr}(g_1^{-1}dg, dg_2g_2^{-1}),$$

$g_1, g_2 \in G^D$, $c \in \mathbf{C}$, and tr will always denote the trace in the adjoint representation. Then *there exists a remarkable imbedding of the gauge group \mathcal{G} into \hat{G}^D as a normal subgroup and $\hat{G} = \hat{G}^D/\mathcal{G}$* . The existence of this imbedding is based on the fact that for any $g \in C^\infty(S^3, G)$, $(\delta^2/12\pi) \int_{S^3} \text{tr}(g^{-1}dg)^3$ determines the class of g in $\pi_3(G) \cong \mathbf{Z}$. This implies that for $g \in \mathcal{G}$ and any extension $\bar{g} \in C^\infty(D^3, G)$, where D^3 the 3-dim ball, $e^{il\omega(g)}$ with $\omega(g) = (\delta^2/12\pi) \int_{D^3} \text{tr}(\bar{g}^{-1}d\bar{g})^3$ is well defined. Then one can check that $\mathcal{G} \ni g \rightarrow (g, e^{il\omega(g)}) \in \hat{G}^D$ is the desired imbedding. Note that the topology of G implies the integrality condition $l \in \mathbf{Z}$, on the central extension \hat{G} . We call l the level of the extension and the central extension of level 1, the basic extension. This notion corresponds to the notion of level l representations of the affine Lie algebra $\hat{\mathfrak{g}}$.

4.2. Characters and analysis on affine Lie algebras. The study of the formal analysis on the affine Cartan subalgebra was launched in [Macdonald (a)]. In the finite-dimensional case the ring $\mathbf{C}[P]^{\pm W}$ of W -(anti-)invariant polynomials on the weight lattice P has a natural realization in terms of Fourier series. The

affine analogue of this ring was correctly interpreted as the ring of elliptic theta-functions in [Macdonald (a)], [Looijenga (a)]. A typical element of $\mathbf{C}[\hat{P}]^{\hat{W}}$ is $\sum_{w \in W} \theta_{w\hat{\lambda}}(h, q)$, $\hat{\lambda} = \lambda + ld \in \hat{P}$, where

$$\theta_{\hat{\lambda}}(h, q) = \sum_{\gamma \in Q^\vee} e^{i\langle \lambda + l\gamma, h \rangle} q^{\langle \lambda, \gamma \rangle + \langle \gamma, \gamma \rangle / 2}$$

is the theta-function. I. Macdonald produced a formal proof of the analogue of the classical Weyl identity for the minimal element $j(h) \in \mathbf{C}[P]^{-W}$ identifying one special case as the Jacobi triple product identity. Later simpler proofs of Macdonald's identity were given, based on the special properties of elliptic functions, namely, a proof using the heat equation on the torus was given in [Kostant], [7], and another using the simple behavior with respect to the modular group $\mathrm{PSL}_2(\mathbf{Z})$ [Van Asch], [Looijenga (a)].

Identities and other structures on the Cartan subalgebra usually reflect some deeper facts for the Lie algebras and groups. In particular, new interpretations and proofs of the Macdonald identity were given in terms of affine Lie algebras [Kac (g)], [Garland and Lepowsky], [Garland (a)].

Now we consider analysis on affine Lie algebras and its relation to the ring $\mathbf{C}[\hat{P}]^{\pm \hat{W}}$. We will follow [Frenkel (g)]. According to the Kirillov program (see, e.g., [Kirillov '78]) the integrals of the one-dimensional character $e^{i\langle x, y \rangle}$ over the coadjoint orbits provide the characters of unitary representations. In the case of the finite-dimensional compact simple Lie algebra \mathfrak{g} and an irreducible representation V_λ one has

$$\mathrm{tr}(\exp ix)|_{V_\lambda} = \frac{1}{p(x)} \int_{\mathcal{O}_{\lambda+\rho}} e^{i\langle x, y \rangle} d\mu(y), \quad ix \in \mathfrak{g}, \quad (3)$$

where μ is the measure on the orbit $\mathcal{O}_{\lambda+\rho}$ induced by Lebesgue measure in the Euclidean space \mathfrak{g} , $\rho = \frac{1}{2} \sum_{\alpha > 0} \alpha$, and $p(x)$ is a fixed function. When $x \in \mathfrak{h}$, the integral reduces to the Weyl character formula. The formal extension of (3) to the affine case indicates that the correct analogue of Lebesgue measure is Gaussian measure. We have seen in §4.1 that the coadjoint orbit in the affine case has a realization in terms of the path space on G and the correct measure on the orbit turns out to be a conditional Wiener measure.¹ *The character of $V_{\hat{\lambda}}$ is represented by a path integral with respect to a conditional Wiener measure of variance $t > 0$.*

$$\mathrm{tr}(\exp(ix - td))|_{V_{\hat{\lambda}}} = \frac{1}{\hat{p}(x, e^t)} \int_{\mathrm{Path}(\mathcal{O}(\hat{\lambda}))} \exp\left(i \int_{S^1} \langle x, g^{-1} dg \rangle\right) dw^t(g), \quad (4)$$

where $x \in \bar{\mathfrak{g}}$, and $\hat{p}(x, e^t)$ is a fixed function. The Brownian motion or the Wiener process is governed by the heat equation, which appeared before in the relation to the Macdonald identity, naturally yielding the Kac-Weyl character formula

¹Let $V_{\hat{\lambda}}$ be a standard representation of the affine Lie algebra $\tilde{\mathfrak{g}}$.

for $x \in \mathfrak{h}$ formulated in terms of theta-functions

$$\mathrm{tr}(\exp(ix - td))|_{V_{\hat{\lambda}}} = \frac{1}{\hat{j}(h, e^t)} \sum_{w \in W} \varepsilon(w) \theta_{\hat{\lambda} + \hat{\rho}}(x, e^t),$$

where $\hat{j}, \hat{\rho}$ are the affine analogues of j, ρ , respectively.

4.3. *Regular representations and problems of analysis on affine Lie groups.* The analysis on affine Lie groups is harder than that on affine Lie algebras. It is an open problem to find a correct analogue of Haar measure on a compact group. The solution of this problem can lead to the discovery of a new measure (or its substitute) which is no less canonical than Wiener measure. The main test for the new measure is the generalization of the Peter-Weyl theorem. In the finite-dimensional case we know that we can identify $\coprod_{\lambda} V_{\lambda} \otimes V'_{\lambda}$ (where V_{λ} is an irreducible highest weight representation, V'_{λ} its dual, and λ runs through all the dominant highest weights) with the space $\mathrm{Pol}(G)$ of $G \times G$ finite functions on G via the matrix coefficients $(\pi_{\lambda}(g)v_i, v_j)$, $g \in G$, $v_i, v_j \in V_{\lambda}$. One can recover the Hermitian structure of V_{λ} defining the pre-Hilbert structure on $\mathrm{Pol}(G)$ by means of the Haar measure μ , and obtain the decomposition $L^2(G, \mu) = \bigoplus_{\lambda} V_{\lambda} \otimes V'_{\lambda}$. In the affine case any matrix coefficient, which is a function on \hat{G} , can be lifted to \hat{G}^D . Choosing the natural section $G^D \rightarrow \hat{G}^D$, $g \rightarrow (g, 1)$, we can identify the matrix coefficients of the representations of the fixed level l with the covariant functions on G^D satisfying the following conditions:

$$f(gg_0) = e^{il(\gamma(g, g_0) + \omega(g_0))} f(g), \quad g_0 \in \mathcal{G}, \quad g \in G^D. \quad (5)$$

This provides an imbedding of $W_l = \coprod_{\text{level } \hat{\lambda}=l} V_{\hat{\lambda}} \otimes V'_{\hat{\lambda}}$ into a space of covariant functions on G^D . In order to define a correct measure, one has to “factor-out” the action of the gauge group, a typical procedure in quantum field theory. The problem is to *give a rigorous construction of the analogue of Haar measure, which allows us to reconstruct the Hermitian product in $V_{\hat{\lambda}}$ and to characterize appropriate completion of W_l .*

Finally, we will consider another realization of the level l representation of $\hat{G} \times \hat{G}$. Let \mathcal{A} be the linear space of \mathfrak{g} -valued 1-forms on D . The group G^D acts naturally on \mathcal{A} : $g \cdot A = g^{-1}Ag + g^{-1}dg$, $g \in G^D$, $A \in \mathcal{A}$. We consider the space of functions on $\mathcal{A} \times \mathcal{A}$ satisfying the following invariance condition with respect to the gauge group $\mathcal{G} \times \mathcal{G} \subset \hat{G}^D \times \hat{G}^D$ (cf. [16b])

$$\begin{aligned} \pi_l(g_L, g_R) f(A_L, A_R) &\equiv e^{il\beta(U, A_L, A_R)} f(g_L \cdot A_L, g_R \cdot A_R) = f(A_L, A_R), \\ \beta(U, A_L, A_R) &= \frac{\delta^2}{4\pi} \int_D \mathrm{tr}(A_R U^{-1} A_L U + A_R U^{-1} dU \\ &\quad + U dU^{-1} A_L - A_R A_L) + \omega(U), \end{aligned} \quad (6)$$

where $U = g_L g_R^{-1}$, $(g_L, g_R) \in \mathcal{G} \times \mathcal{G}$, $(A_L, A_R) \in \mathcal{A} \times \mathcal{A}$, and β is the one-cocycle of $\mathcal{G} \times \mathcal{G}$ known in physics literature as the Wess-Zurmino effective action in two dimensions (see, e.g., [17] for a review). The action of the group $\hat{G}^D \times \hat{G}^D$ is defined by the same formula (6) with the subtraction of $\omega(g_L) - \omega(g_R)$ from the cocycle

β , which comes from the imbedding of the gauge group. This results in the representation of the factor group $\hat{G} \times \hat{G}$.

The most remarkable property of the line bundle (6) over $\mathcal{A} \times \mathcal{A}/\mathcal{G} \times \mathcal{G}$ is that for $G = \mathrm{SU}(n)$, $l = 1$, it can be naturally realized as the determinant line bundle for the Dirac operator in two dimensions [17] (in the existing literature \mathcal{A} is defined on S^2 rather than on D ; see also §9). Integer powers of this line bundle provide all higher levels of representation of $\widehat{\mathrm{SU}(n)} \times \widehat{\mathrm{SU}(n)}$. The rigorous construction of the measure in $\mathcal{A} \times \mathcal{A}$, which would provide the Hermitian structure and the characterization of a completion of W_l is an important open problem.

5. Affine Lie algebras as operator algebras. In this section we consider certain generalizations of affine Lie algebras, namely, affine Lie algebras of infinite rank. These algebras are closely related to inductive limits of classical finite-dimensional Lie algebras and admit generalizations of many classical constructions. They also have natural operator realizations, and since they contain affine Lie algebras as subalgebras, provide another approach to our subject. This brings further applications of affine Lie algebras to symmetric groups, KdV equations, Bott's periodicity, etc.

5.1. Operator realization. There are three classes of simply laced affine Lie algebras of infinite rank $A_\infty, D_\infty, E_\infty$ (or $A_{+\infty}$) and two twisted classes. We will consider only the type A_∞ . Let M_∞ be the associative algebra of infinite matrices $x = (x_{ij})_{i,j \in \mathbf{Z}+1/2}$ such that $x_{ij} = 0$ for $|i - j| > N(x)$. Let \mathfrak{G} denote either the complex Lie algebra $\mathfrak{gl}(\infty, \mathbf{C})$ of matrices from M_∞ with the natural Lie bracket or its compact form $\mathfrak{u}(\infty)$, consisting of the skew-Hermitian matrices. Then \mathfrak{G} is the affine Lie algebra of type A_∞ . One has $\dim H^2(\mathfrak{G}) = 1$ and the central extension $\hat{\mathfrak{G}}$ is defined by the two-cocycle $\alpha(x, y) = \mathrm{tr}(x_{+-}y_{-+} - y_{+-}x_{-+})$, where $x_{+-} = (x_{ij})_{i, -j \in (\mathbf{Z}+1/2)_+}$, etc. The subalgebra $\mathfrak{G}_n \subset \mathfrak{G}$ of block periodic matrices $x_{ij} = x_{i+n, j+n}$, $i, j \in \mathbf{Z} + 1/2$, is isomorphic to $\overline{\mathfrak{gl}(n, \mathbf{C})}$ or $\overline{\mathfrak{u}(n)}$. The restriction of the two-cocycle on \mathfrak{G} yields the cocycle α on \mathfrak{G}_n (see §4.1) [Date, Jimbo, Kashiwara and Miwa (a)]. This realization suggests a natural "maximal" completion to the restricted algebra of bounded operators x in the Hilbert space H with the Hilbert-Schmidt $x_{+-} : H_+ \rightarrow H_-$ and $x_{-+} : H_- \rightarrow H_+$, where x is identified with an infinite matrix in an orthonormal basis $\{e_i, i \in \mathbf{Z} + 1/2\}$ of H , and $H_\pm = \bigoplus_{\pm i > 0} \mathbf{C}e_i$.

In the definition of the central extension of the group \mathbf{G} of invertible bounded operators with the Lie algebra \mathfrak{G} we come across the same difficulty as in §4, namely, there is no continuous 2-cocycle on \mathbf{G} . Now we can follow the analogy with the central extension of $\mathrm{SU}(m, n)$. In this case the cocycle can be defined explicitly on the open set $\mathbf{G}_* = \{g \in \mathbf{G} : g_{++} \text{ is invertible}\}$ by the formula $\gamma(g_1, g_2) = \det(a_1 a_2 a_3^{-1})$, where $a_i = g_{i++}$ and $g_3 = g_1 g_2$. This cocycle still makes sense in the ∞ -dimensional case and corresponds to the Lie algebra cocycle. In order to define the group $\hat{\mathbf{G}}$ we "resolve" this cocycle following [Pressley and Segal]. Let $\mathbf{G}^2 = \{(g, q) \in \mathbf{G} \times \mathbf{G}_{++} : a - q \text{ of trace class}\}$, and

let $\mathcal{G} = \{q \in \mathbf{G}_{++} : \det q = 1\}$. Then \mathcal{G} is a normal subgroup of \mathbf{G}^2 under the imbedding $q \rightarrow (\text{Id}, q)$ and $\hat{\mathbf{G}} = \mathbf{G}^2/\mathcal{G}$ is the central extension corresponding to $\hat{\mathfrak{G}}$ (cf. §4). Note that the choice of the section $(g, a) \in \mathbf{G}^2$ corresponding to \mathbf{G}_* defines the cocycle γ . The restriction of this construction to the block diagonal subgroup yields the basic extension $\widehat{\text{GL}}(n)$ or $\hat{\text{U}}(n)$.

5.2. Determinant bundle and spin representation. Let $\hat{\mathbf{G}}_0$ be the identity component of $\hat{\mathbf{G}}$. We construct its fundamental representation in the determinant bundle again following the analogy with the finite-dimensional case. We consider the Stiefel manifold bundle

$$\text{St}(H^+, H) = \{w: H_+ \rightarrow H; w_+ \text{ has a determinant, } w_- \text{-Hilbert-Schmidt}\}$$

over the corresponding Grassmann manifold $\text{Gr}(H_+, H)$. The identification of bases in $\text{St}(H_+, H)$, which differ only by \mathcal{G} , provides a determinant line bundle over $\text{Gr}(H_+, H)$. The group $\hat{\mathbf{G}}_0$ acts naturally on this bundle via $(g, q) \cdot w = gwq^{-1}$. The determinant line bundle, its Hermitian and holomorphic structure, and the representations of $\hat{\mathbf{G}}_0$ have been studied thoroughly in [Pressley and Segal].

Another construction of the fundamental representations of $\hat{\mathfrak{G}}$, the spin representation, is again suggested by the finite-dimensional theory. We let H^* denote the Hermitian dual of H . Let $C(H \oplus H^*)$ be the Clifford algebra generated by e_n, e_m^* , $n, m \in \mathbf{Z} + 1/2$, $\{e_n, e_m^*\} = \delta_{n,m}$, and let $V = \Lambda(H_- \oplus H_+^*)$ be the Clifford module. We define the representation of the elementary matrix $E_{m,n}$ as $e_m e_n^* - \delta_{m,n} \theta(m)$, where $\theta(m) = 1$, $m > 0$, and 0 , $m < 0$. One can check that the infinite linear sums of operators $\sum_{m,n} x_{m,n} E_{m,n}$, $x = (x_{m,n}) \in \mathfrak{G}$, are well defined and we obtain the representation of $\hat{\mathfrak{G}}$ with the correct cocycle α . As in the finite-dimensional case the determinant line bundle representation is equivalent to the spin representation [Pressley and Segal].

The representation V has a natural decomposition, with respect to $\hat{\mathfrak{G}}$, into irreducible representations $V_n = \{e_{n_1} \cdots e_{n_k} e_{m_1}^* \cdots e_{m_l}^*, n = k - l\}$, $n \in \mathbf{Z}$. The remarkable property of these representations is that they stay irreducible for any subalgebra $\hat{\mathfrak{G}}_n$, $n = 1, 2, \dots$ (cf. §6.2). Let us consider $V_0 = \bigoplus_{n \geq 0} V_0^n$ with the grading defined by $\deg e_n = -n$, $\deg e_n^* = n$, $\deg 1 = 0$; then $\dim V_0^n = p(n)$, where p is the partition function. Since $V_0 \cong S(\mathfrak{G}_1^-)$ it has two canonical bases respecting the grading. The table of their inner products for V_0^n , which is the square matrix of order $p(n)$, provides the character table for the symmetric group S_n [Sato '83, unpublished]. The three constructions of the representations V provide a beautiful description of the hierarchies of Korteweg-de Vries equations, reviewed in [Date, Jimbo, Kashiwara, and Miwa (a), (c), (e)] (see also [Segal and Wilson] and [Frenkel (e)]).

Finally, we note that the spin representation of $\hat{\mathfrak{G}}_n$ allows also loop algebra interpretation. In fact, $\hat{\mathfrak{G}}_n \cong \hat{\mathfrak{g}}_n$ for $\mathfrak{g}_n = \mathfrak{gl}(n, \mathbf{C})$ or $\mathfrak{u}(n)$, and similar $H \cong \overline{H}_n = H_n \otimes t^{1/2} \mathbf{C}[t, t^{-1}]$, $e_i + mn \leftrightarrow e_{i+1/2} \otimes_n t^{m+1/2}$, $m \in \mathbf{Z}$, $i + 1/2 = 1, \dots, n$, the space of antiperiodic loops in $H_n = \bigoplus_{i=1}^n \mathbf{C} e_i$. Then the above construction of the spin representation of $\hat{\mathfrak{g}}$ is analogous to the classical spin representation

of \mathfrak{g}_n in $\Lambda(H_n^*)$ and was known to physicists in the context of quark models [2]. The natural change of polarization $\Lambda(H^*)$ to $\Lambda(H_- \oplus H_+^*)$ in the affine case was first suggested by P. Dirac in quantum theory.

The spin representation of $\hat{\mathfrak{g}}_n$ in V can be extended to the affine Lie algebra $\hat{\mathfrak{o}}(2n)$ and is decomposed into two irreducible representations. If we choose \bar{H}_n to be the space of periodic loops in H_n , then a slight modification of the above construction provides the other two fundamental level 1 representations of $\hat{\mathfrak{o}}(2n)$ [Frenkel (b), (d)], [Kac and Peterson (b)]. The appropriate analogue of the determinant bundle in the orthogonal case is the pfaffian bundle [Pressley and Segal]. Generalization of the spin representation to the metaplectic representation and other classical affine algebras can be found in [Feingold and Frenkel (b)].

6. Affine Lie algebras as vertex algebras. Level 1 representations of affine Lie algebras and groups of ADE type have a remarkable property which has no direct analogue in the finite-dimensional case: they remain irreducible under the restriction to various Heisenberg subgroups, and therefore have a particularly simple structure. The reconstruction of the rest of the affine Lie algebra is given in terms of the so-called vertex operators, which represent “ δ -loops,” or strictly speaking, some generating functions. We will discuss applications of vertex operators to various physical models.

6.1. Vertex representation. Every simply laced affine Lie algebra possesses at least one basic representation, namely, the fundamental representation V corresponding to the additional point of the extended Dynkin diagram. The study of basic representations elucidates the unitary and integral structure in the affine, as well as in the finite-dimensional case. We will follow [Frenkel and Kac] and [Segal].

Let $\mathfrak{g}_{\mathbb{C}}$ be a simple finite-dimensional Lie algebra of ADE type, with Cartan subalgebra $\mathfrak{h}_{\mathbb{C}}$, root system Δ , basis Π , root lattice Q , and weight lattice P . We will normalize the invariant symmetric bilinear form $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}_{\mathbb{C}}$ so that $\langle \alpha, \alpha \rangle = 2$, $\alpha \in \Delta$. We define a bilinear two cocycle $\varepsilon: Q \times Q \rightarrow \{\pm 1\}$ such that $\varepsilon(\alpha, \alpha) = (-1)^{\langle \alpha, \alpha \rangle / 2}$, $\alpha \in Q$. Then $\mathfrak{g}_{\mathbb{C}}$ can be constructed explicitly in terms of its Chevalley basis $\Pi = \{\alpha_i\}$, x_{α} , $\alpha \in \Delta$ equipped with the cocycle ε :

$$\begin{aligned} [\alpha_i, \alpha_j] &= 0, & [\alpha_i, x_{\alpha}] &= \langle \alpha_i, \alpha \rangle x_{\alpha}, & [x_{\alpha}, x_{-\alpha}] &= -\alpha, \\ [x_{\alpha}, x_{\beta}] &= \varepsilon(\alpha, \beta) x_{\alpha+\beta}, & \alpha + \beta &\in \Delta; & \text{and } 0, \alpha + \beta &\notin \Delta \cup 0. \end{aligned} \quad (7)$$

This provides, in particular, the simplest construction of the Lie algebra of type E_8 . The construction (including the Chevalley basis) for $\hat{\mathfrak{g}}_{\mathbb{C}}$ is then straightforward.

According to the Stone–von Neumann theorem the Heisenberg subalgebra $\hat{\mathfrak{h}}_{\mathbb{C}}$ has an essentially unique representation in the Fock space $S(\mathfrak{h}_{\mathbb{C}}^-)$ via multiplication and derivation operators. The remarkable property of the basic representation V is that $V|_{\hat{\mathfrak{h}}} \cong \bigoplus_{\gamma \in Q} S(\mathfrak{h}_{\mathbb{C}}^-)^{\gamma} \cong S(\mathfrak{h}^-) \otimes \mathbb{C}[Q]$, where $\mathbb{C}[Q]$ is the group algebra of Q , twisted by ε , i.e., $e^{\alpha} \cdot e^{\beta} = \varepsilon(\alpha, \beta) e^{\alpha+\beta}$, $\alpha, \beta \in Q$. The

reconstruction of the whole algebra $\hat{\mathfrak{g}}_{\mathbf{C}}$ is given in terms of the vertex operators

$$\begin{aligned} X(\alpha, z) &= : \exp q_{\alpha}(z) : , \quad \alpha \in Q, \quad z \in \mathbf{C} \setminus 0, \\ q_{\alpha}(z) &= \sum_{n \neq 0} \alpha(-n) z^n / n + \alpha(0) \log z + \alpha, \end{aligned} \quad (8)$$

where $:$ denotes the normal ordering, i.e., the rearranging of all terms e^{α} , $\alpha(n)$, $n \in \mathbf{Z}$, so that the “creation operators” e^{α} , $\alpha(n)$, $n < 0$, occur to the left of the “annihilation operators” $\alpha(n)$, $n \geq 0$. Then the homogeneous components $X_n(\alpha)$ of $X(\alpha, z) = \sum_{n \in \mathbf{Z}} X_n(\alpha) z^{-n-1}$ are well defined operators in V and, for $\alpha \in \Delta$, provide the representation of $x_{\alpha} \otimes t^n \in \hat{\mathfrak{g}}$. Note that $q_{\alpha}(z)$ is just a formal expression, though $p_{\alpha}(z) = dq_{\alpha}(z)/dz = \sum_{n \in \mathbf{Z}} \alpha(-n) z^{n-1}$ is the generating function for the Heisenberg subalgebra.

The vertex construction intrinsically provides V , $\hat{\mathfrak{g}}$, and $\mathcal{U}(\hat{\mathfrak{g}})$ with unitary and integral structure (see [Mitzman (b)] for details), studied before in [Garland (b), (c)]. Also, the character of V (found in [Feingold and Lepowsky] for \hat{A}_1 and [Kac (j)] in general) displays a modular structure since

$$\chi(V) \equiv \text{tr}(\exp 2\pi i(h + \tau d))|_V = \frac{\theta_Q(h, q)}{\eta(q)^r},$$

where $r = \text{rank } Q$, $\theta_Q(h, q) = \sum_{\gamma \in Q} e^{2\pi i \langle \gamma, h \rangle} q^{\langle \gamma, \gamma \rangle / 2}$, $q = \exp 2\pi i \tau$, $\text{Im } \tau > 0$, is invariant with respect to a congruence subgroup of the modular group $\text{PSL}_2(\mathbf{Z})$.

A slight modification of the vertex construction, namely, the change of Q to a Q -coset in P , provides all level 1 fundamental representations of $\hat{\mathfrak{g}}_{\mathbf{C}}$. Restriction of the level 1 representations to the subalgebra $\hat{\mathfrak{g}}_{\mathbf{C}}[n] = \mathfrak{g}_{\mathbf{C}} \otimes \mathbf{C}[t^n, t^{-n}] \oplus \mathbf{C}c \cong \hat{\mathfrak{g}}_{\mathbf{C}}$ provides the construction of all the dominant highest weight representations. The multiplicities of V_{λ} turn out to be given by $\chi_q(V_{\lambda})/\chi_q(V)$, where χ_q is the principally specialized character. This implies that the sum

$$\sum_{\lambda \text{ of level } l} \chi(V_{\lambda}) \chi_q(V_{\lambda}) = \frac{\theta_P(h, q)}{\eta(q)^r} \chi_q(V)$$

does not depend on the level [Frenkel (f)]. The detailed study of the modular behavior of the characters $\chi(V_{\lambda})$ was done in [Kac and Peterson (f)], where the relation with Hecke modular forms was found.

There exist twisted versions of the vertex construction. In particular, the twisted vertex representation of $\hat{\mathfrak{sl}}(2)$ [Lepowsky and Wilson (a)] was the first construction of this sort in the mathematical literature. Another twisted construction [Frenkel, Lepowsky, and Meurman (a)] plays an important role in the construction of the Monster (cf. §8.3). The twisted construction of [Kac, Kazhdan, Lepowsky, and Wilson] underlies the structure of the KdV hierarchies [Date, Jimbo, Kashiwara, and Miwa (a)] (cf. §5.2). The most general treatment of twisted vertex representations was given in [Lepowsky (j)] and [Kac and Peterson (i)]. The detailed study of the higher level twisted representations and their relation to generalizations of the Rogers-Ramanujan identities was done in [Lepowsky and Wilson (e), (g)] (see also [Meurman and Primc]).

Some components of the vertex construction were found earlier by physicists (particularly see [13]) in the context of the string theory. Recently, the untwisted vertex representation was used in the heterotic string model [12], which has been proposed as a candidate for a unified theory, including gravity.

6.2. Bose-Fermi correspondence. We obtained in §§5.2 and 6.1 two explicit realizations of all four level 1 fundamental representations of the affine Lie algebra $\hat{\mathfrak{o}}(2n)$. Their isomorphism can be given in terms of vertex operators [Frenkel (d)]. This result was essentially known in two-dimensional quantum field theory as the Bose-Fermi correspondence (see [13] and references therein). Let us consider the vertex representation V choosing the lattice $\mathbf{Z}^n \subset P$. We define $\varepsilon : \mathbf{Z}^n \times \mathbf{Z}^n \rightarrow \{\pm 1\}$ to be bilinear and to satisfy $\varepsilon(\alpha, \beta)\varepsilon(\beta, \alpha) = (-1)^{\langle \alpha, \beta \rangle + \langle \alpha, \alpha \rangle \langle \beta, \beta \rangle}$. Consider now the vertex operators $X(\alpha, z)$ corresponding to the unit vectors $\alpha = \pm e_i$, $i = 1, \dots, n$, of \mathbf{Z}^n . Then the only nontrivial anticommutators are

$$\{X_m(e_i), X_{-m}(-e_i)\} = 1, \quad i = 1, \dots, n, \quad m \in \mathbf{Z} + 1/2.$$

Thus $X_m(\pm e_i)$ can be identified with the generators $e_i \otimes t^m$, $(e_i \otimes t^{-m})^*$ of the Clifford algebra $C(\overline{H}_n \oplus \overline{H}_n^*)$. Also, it is easy to show that $X_m(\pm e_i)1 = 0$, $m > 0$. Thus the Clifford module $\Lambda(\overline{H}_n \oplus \overline{H}_n^{+*})$ is identified with a subspace of the vertex representation V and in fact, they coincide since V has the same decomposition into two irreducible representations of $\hat{\mathfrak{o}}(2n)$ yielding again the Jacobi triple product identity. The Bose-Fermi correspondence can be generalized for two other representations of $\hat{\mathfrak{o}}(2n)$ as well as for the twisted orthogonal Lie algebra $\hat{\mathfrak{o}}^{(2)}(2n+2)$ [Frenkel (d)]. Thus we have brought into use the vertex operators $X(\alpha, z)$ with $\langle \alpha, \alpha \rangle = 1$ in addition to the ones with $\langle \alpha, \alpha \rangle = 2$ of §6.1. We will consider more general vertex operators in §8.

7. Virasoro algebra. Starting from the definition of an affine Lie algebra, one comes across its everpresent companion, the Virasoro algebra $\hat{\mathcal{L}}$, but the most important relation between them is due to the fact that the Virasoro algebra acts on an affine Lie algebra as a derivation algebra. We will see another relation in §8.2. In the loop algebra realization $\hat{\mathcal{L}}$ is naturally realized as the central extension of the algebra of vector fields on the circle, spanned by $L_n = e^{ins}(1/i)d/ds$, $n \in \mathbf{Z}$. The coadjoint action of the corresponding group yields the Schwarzian derivative as the one cocycle, and orbits of this action have been studied before in the theory of Hill operators [Segal].

The representation theory of this algebra is no less remarkable. A great deal of information about the structure of irreducible highest weight representations is contained in the determinant formula [Kac '78], [Feigin and Fuks (b), (c)], [Rocha-Caridi and Wallach (c), (d)]. Recently, in the relation with the models of statistical mechanics, the discrete series of unitary representations with the values of $c = 1 - 6/(m+1)(m+2)$, $m = 1, 2, 3, \dots$, was discovered in [Friedan, Qui, and Shenker], and their explicit construction was given in [Goddard, Kent, and Olive]. The characters of these unitary representations have been found in [Rocha-Caridi (d)], and they turn out to be modular functions with respect to

congruence subgroups of $\mathrm{PSL}_2(\mathbf{Z})$. We remarked that in the affine case unitary, integral, and modular structures of representations mysteriously invoke each other. In the case of the discrete series of the Virasoro algebra we know only two of the three constituents. *The construction of the discrete series representations of the group corresponding to $\hat{\mathcal{L}}$* is an interesting open problem.

8. Generalization of affine Lie algebras via vertex construction and critical dimension. The realization of affine Lie algebras as vertex algebras suggests some natural generalizations of the latter, which allow us to obtain, in particular, some new information about Kac-Moody algebras. These generalized vertex algebras behave especially simply in rank 24 or its counterpart (25, 1). This critical number is intrinsic to the structure of finite-dimensional simple Lie algebras ("strange formula" of Freudenthal-de Vries), even integral lattices (Leech lattice), modular forms (24th power of η -function), the Virasoro algebra (Weyl's reflection), and as we will see, the vertex algebras. All these structures turn out to be interrelated. The culmination of this phenomenon is the appearance of the Monster group and its remarkable representation.

8.1. Vertex algebras and Kac-Moody algebras. The vertex representation of affine Lie algebras suggests a natural generalization. Let L be an even integer lattice (not necessarily positive definite) of rank r , and let $\varepsilon: L \times L \rightarrow \{\pm 1\}$ be a bilinear function such that $\varepsilon(\alpha, \alpha) = (-1)^{\langle \alpha, \alpha \rangle / 2}$, $\alpha \in L$. We define V as follows (cf. §6): Let $\mathfrak{h} = L \otimes \mathbf{C}$, $\mathfrak{h}^- = \mathfrak{h} \otimes t^{-1}\mathbf{C}[t^{-1}]$; then $V = S(\mathfrak{h}^-) \otimes \mathbf{C}[L]$. V is Hermitian iff L is positive-definite. Then, for any $\alpha \in L$, one can define the vertex operator $X(\alpha, z)$ as above (8). In particular, if L is the root lattice of a simply laced Kac-Moody algebra, then the zero components $X_0(\alpha)$, $\alpha \in \Delta_R$, generate this algebra, thus providing an imbedding of the Kac-Moody algebra into an associative algebra of generalized vertex operators [Frenkel (f)], [Goddard-Olive (a)]. To describe a closed Lie algebra containing the Kac-Moody algebra, we attach a generalized vertex operator $X(v, z)$ to any $v \in V$ following [3]. In particular,

$$X(e^\alpha, z) \equiv X(\alpha, z), \quad X(h(-n), z) \equiv \frac{1}{(n-1)!} \left(\frac{d}{dz} \right)^{n-1} h(z),$$

and to the product of $h_i(-n_i)$ and e^α we assign the normal ordered product of the corresponding vertex operators. Then $X_0(v) \equiv 0$ iff $v \in L_{-1}V$, $L_{-1} \in \hat{\mathcal{L}}$, and the factor space $\mathcal{V} = V/L_{-1}V$ is closed with respect to the Lie bracket. Moreover, *the natural action of \mathcal{V} on itself coincides with the adjoint action* $[X_0(u), X_0(v)] = X_0(X_0(u)v)$, $u, v \in \mathcal{V}$. Let $V^n = \{v \in V : (L_m - n\delta_{0,m})v = 0, m \geq 0\}$; then $\mathcal{V}^1 = V^1/L_{-1}V^0$ is a subalgebra of \mathcal{V} containing the Kac-Moody algebra. The multiplicities of \mathcal{V}^1 yield an upper bound for the multiplicities of the Kac-Moody algebra: $\mathrm{mult}(\mathcal{V}^1)^\alpha = p_{r-1}(1 - \langle \alpha, \alpha \rangle / 2) - p_{r-1}(-\langle \alpha, \alpha \rangle / 2)$, where p_{r-1} is the number of partitions into $r-1$ colors. These multiplicities were found first in string theory [4]. In the next subsection we will obtain the exact upper bound, depending on $\langle \alpha, \alpha \rangle$, for the multiplicities of the Kac-Moody algebra in the critical dimension.

8.2. No-ghost theorem and cohomology of Virasoro algebra. The subspace V^1 is known in string theory as the physical subspace V^{phys} . In the critical dimension, i.e., when $\text{signature}(L) = (25, 1)$, it has a remarkable unitary structure given by the so-called no-ghost theorem, [11], [4]. For any $\alpha \in L$, $\alpha(-1) \in V^1$, we fix a vector of negative norm $\alpha_0 \in L$. The no-ghost theorem asserts that *the orthogonal complement of $\alpha_0(-1)$ in V^1 is positive-semidefinite*. The radical of the Hermitian form contains $L_{-1}V^0$ and is an ideal in \mathcal{V}^1 . Then the factor algebra $\mathcal{V}_0^1 = V^1/\text{Radical}$ still contains the Kac-Moody algebra and $\text{mult}(\mathcal{V}_0^1)^\alpha = p^{(24)}(1 - \langle \alpha, \alpha \rangle/2)$ [Frenkel (f)], [3].

A conceptual interpretation and proof of the no-ghost theorem are given in [8] (cf. [14] in physics literature) and are based on the semi-infinite cohomology defined originally in [Feigin]. Let $\mathfrak{G} = \coprod_{n \in \mathbb{Z}} \mathfrak{G}_n$ be a graded Lie algebra such that $\dim \mathfrak{G}_n < +\infty$ for each n , and let the dual algebra $\mathfrak{G}' = \coprod_{n \in \mathbb{Z}} \mathfrak{G}'_n$. We fix $\{e'_i\}_{i \in \mathbb{Z}}$, a homogeneous basis of \mathfrak{G}' , so that if $e'_i \in \mathfrak{G}'_n$ then $e'_{i+1} \in \mathfrak{G}'_n$ or \mathfrak{G}'_{n+1} . We define the space of semi-infinite forms $\Lambda_\infty^* \mathfrak{G}'$ spanned by the monomials $e'_{i_1} \wedge e'_{i_2} \wedge \cdots$ such that for sufficiently large k , $i_k + 1 = i_{k+1}$. Then if $H^2(\mathfrak{G}) = 0$ one can define the semi-infinite cohomology $H_\infty^*(\mathfrak{G}, \mathbf{V})$ with coefficients in a graded \mathfrak{G} -module $\mathbf{V} = \coprod_{n \in \mathbb{Z}_+} \mathbf{V}_n$, and corresponding relative cohomology with respect to a subalgebra of \mathfrak{G}_0 . In the case when $\hat{\mathfrak{G}}$ is the Virasoro algebra $\hat{\mathcal{L}}(2)$ and $\mathbf{V} = V$, $H_\infty^*(\hat{\mathcal{L}}, Cc, V)$ is trivial unless $\text{rank } L = 26$, which singles out the critical dimension. When $\text{signature}(L) = (25, 1)$,

$$H_\infty^0(\hat{\mathcal{L}}, \hat{\mathcal{L}}_0, V) \cong H_\infty^{\pm 1/2}(\hat{\mathcal{L}}, Cc, V) \cong V^{\text{phys}}/\text{Radical}, \quad (10)$$

thus giving a cohomological interpretation of the vector space of the vertex algebra \mathcal{V}_0^1 . It is natural to anticipate that this Lie algebra is induced from a larger Lie algebra on the complex $\Lambda_\infty^*(\mathcal{L}) \otimes V$. In [20b] overwhelming evidence was presented that this larger Lie algebra has a geometrical realization in terms of interacting strings. Further *study of the isomorphism (10) with the additional Lie algebra structure* is of indubitable interest.

8.3. Monster and Moonshine. The existence and the construction of the biggest sporadic (finite simple) group F_1 is related in a remarkable way to the vertex algebras in the critical dimension. This sporadic group, called the Monster group, was first discovered by B. Fischer and R. Griess and later constructed by Griess (see [Griess '83] for a review). In [Griess (b)] he constructs the smallest nontrivial irreducible complex representation B_0 of F_1 , of dimension 196883, having the property that $S^2(B_0)$ contains B_0 and the trivial representation \mathbf{C} with multiplicity 1. This property implies the existence of an F_1 -invariant commutative algebra structure in B_0 , $(x, y) \mapsto xy$ as well as a bilinear symmetric invariant form $\langle x, y \rangle$, $x, y \in B_0$. B_0 equipped with both structures is called the Griess algebra. One can adjoin the identity e , to get $B = B_0 \oplus \mathbf{C}e$, defining the new product \times by $x \times y = xy + \langle x, y \rangle e$, $x, y \in B_0$. J. McKay observed that the dimension of B_0 differs from the smallest nontrivial coefficient of the normalized modular invariant $J(q) = q^{-1} + 0 + 196884q + \cdots$ by only 1. Similar properties of the next coefficients led to the conjecture [Thompson] that there

is a natural graded representation of F_1 , $V^{\mathfrak{h}} = \bigoplus_{n \geq -1} V_n^{\mathfrak{h}}$ with the weighted character $\chi_q(V^{\mathfrak{h}}) = \sum_{n \geq -1} \dim V_n^{\mathfrak{h}} q^n$ equal to $J(q)$. This representation was explicitly constructed in [Frenkel, Lepowsky, and Meurman (b), (c)], [9] using the representation theory of affine Lie algebras and the technique of generalized vertex operators. Furthermore, we constructed in $V^{\mathfrak{h}}$ a representation π of an “affinization” of the Griess algebra with identity $\hat{B} = B \otimes \mathbb{C}[t, t^{-1}] \oplus \mathbb{C}c$, where the product in \hat{B} is defined by

$$x(m) \times y(n) = (x \times y)(m+n) + m^2 \langle x, y \rangle \delta_{m+n,0} c \quad (11)$$

where $x, y \in B$, $m, n \in \mathbb{Z}$, $x(m) \equiv x \otimes t^m$. The representation π of \hat{B} in the space $V^{\mathfrak{h}}$, with the character given by the modular invariant $J(q)$, is understood in terms of the identity

$$\pi((x \times y)(m+n)) = \frac{1}{2} \{ [\pi(x(m+1)), \pi(y(n-1))] + [\pi(y(n+1)), \pi(x(m-1))] \}. \quad (12)$$

Then the representation of \hat{B} in $V^{\mathfrak{h}}$ is irreducible and is compatible with the action of the Monster group F_1

$$g\pi(x(m))g^{-1} = \pi((g \cdot x)(m)), \quad g \in F_1. \quad (13)$$

It should be remarked that in our approach we construct $V^{\mathfrak{h}}$ and the representation of \hat{B} and F_1 from the very beginning, with only minor use of finite group theory. For the concrete model of $V^{\mathfrak{h}}$ we use the untwisted V and twisted V' vertex representations associated with the Leech lattice. Each of the two vertex representations is decomposed into two subspaces V^{\pm}, V'^{\pm} with respect to the canonical involution σ . We define $V^{\mathfrak{h}} = V^+ \oplus V'^-$ and check that $\chi_q(V^{\mathfrak{h}}) = J(q)$. The construction of \hat{B} and F_1 is based essentially on the triality principle which allows us to intertwine sectors of the nontwisted and the twisted representation. An important open problem is to provide a more invariant description of the above representation \hat{B} and F_1 in the space $V^{\mathfrak{h}}$, presumably in the sense of §8.1.

We have seen that in the critical dimension there exist the two most remarkable unitary and discrete structures, namely, the no-ghost theorem and the representation of the Monster. According to the principle confirmed for affine Lie algebras (§6.1) and the Virasoro algebra (§7), we can expect very rich modular structure. A most remarkable modular structure of $V^{\mathfrak{h}}$ was discovered in [Conway and Norton] and is known as Monstrous Moonshine. The first property that we assumed about $V^{\mathfrak{h}}$ is that its character, which is also the q -trace of the identity element of the Monster, is given by the modular invariant $J(q)$. It was conjectured that the q -trace of any element of the Monster yields a Hauptmodul of the genus zero surface (\widehat{H}/Γ) associated to a discrete subgroup Γ of $\mathrm{PSL}_2(\mathbb{R})$. The construction of the F_1 representation in $V^{\mathfrak{h}}$ [Frenkel, Lepowsky, and Meurman (b), (c)], [9] allows us to find traces for some (and probably for all) elements of the Monster and to check the required properties. However, the complete conceptual explanation of Monstrous Moonshine is still an open problem.

Finally, we remark that the modular behavior of the partition function of $V^{\mathrm{phys}}/\mathrm{Rad}$ in the critical dimension is responsible for the consistency of string

theory at the one-loop level and thus presumably for higher loops (for a review see, e.g., [Schwarz]). Modular invariance is crucial for anomaly cancellations in the heterotic string [12].

9. Generalization of affine Lie algebras via geometry: Current algebras and groups. The realization of affine Lie algebras as loop algebras (see §4) suggests a replacement of S^1 by a more general manifold. We will see in this section that some structures related to affine Lie algebras and groups have perfect higher-dimensional analogues, while some others are not yet known.

Let M be a compact Riemannian manifold, let \mathfrak{g} be a finite-dimensional compact simple Lie algebra, and let G be the corresponding group. We define \mathfrak{g}^M (resp. G^M) as the space of smooth maps from M into \mathfrak{g} (resp. G) with pointwise multiplication. The results about the representation theory of these algebras and groups have been reviewed in [Ismagilov '83]. However, the analogy with affine Lie algebras suggests that the most interesting phenomena occur for their extensions. The crucial idea is that *when $d > 1$, one should look for noncentral extensions*. This has both physical [6], [16], and pure mathematical (joint work with I. Singer, see [19, §9] for the announcement) motivations. Here we only mention the simplest one, and we set $M = S^d$. Note first that $H^2(\overline{G}; \mathbf{R})$ is one-dimensional, and its generator α can be chosen to be left invariant. The loop algebra $\overline{\mathfrak{g}}$ can be identified with the left invariant vector fields on \overline{G} , and α thus defines a central extension. If $d > 1$ then there are no left invariant elements in $H^2(G^M; \mathbf{R})$; however, the second cohomology group can be nontrivial for odd d , and one can define a noncentral extension of \mathfrak{g}^M in the abelian algebra of real functions on G^M . For example, if $d+2 = 2m_i + 1$ for some exponent m_i of \mathfrak{g} , then $H^2(G^M; \mathbf{R}) \neq 0$, which entails a nontrivial extension of \mathfrak{g}^M . The two-cocycle in the case $G = \mathrm{SU}(n)$ can be obtained in an explicit form from families of Dirac operators [19].

Another approach to the explicit form of the 2-cocycle is given by the Chern-Weil theory. Let I be a symmetric G -invariant polynomial on \mathfrak{g} of degree m such that $d = 2m - 3$. Let \mathcal{A} be the space of \mathfrak{g} -valued 1-forms on the ball D^{d+2} . One can define the Chern-Simons form $\omega_0(A)$, $A \in \mathcal{A}$, such that

$$d\omega_0(A) = I(\underbrace{F, \dots, F}_m), \quad F = dA + A \wedge A.$$

Then the two-cocycle on \mathfrak{g}^M , $d > 1$, is obtained from the Chern-Simons form by the descent method [21]:

$$\alpha(x, y; A) = c \int_{S^d} \left\{ \int_0^1 I(dx, dy, A, \underbrace{F_t, \dots, F_t}_{m-3}) (1-t)^2 dt \right\}, \quad (13)$$

where $F_t = t dA + t^2 A \wedge A$, c is a constant. In the case when $\mathfrak{g} = \mathfrak{su}(n)$, $n > 2$, α has an especially simple form

$$\alpha(x, y; A) = c \int_{S^3} \mathrm{tr}([dx, dy]A). \quad (14)$$

We note that α is expressed as an integral over S^d . In fact, one can define a natural notion of local cohomology. It was shown in [5] that the elements of the second cohomology group are in one-to-one correspondence with $S(\mathfrak{g})^G$.

Following the ideas of [16b] one can also construct the corresponding extension of the group G^M . As in the two-dimensional case we consider S^d as the boundary of the ball $D = D^{d+1}$. We set $\omega(g) = c \int_D \omega_0(g^{-1} dg)$. Let $G^D = C^\infty(D^{d+1}, G)$ and $\mathcal{G} = \{g \in G^D : g|_M = \text{Id}\}$ be a gauge group. Then, again starting from the Chern-Simons form, there is a way to construct the 1-cocycle on the group \mathcal{G} , $\beta(g; A) = \beta_0(g; A) + \omega(g)$ [21]. If $g_1, g_2 \in G^D$ then the coboundary $\delta\beta \neq 0$; however $\gamma_0(g_1, g_2; A) \equiv (\delta\beta)(g_1, g_2; A)$ depends only on boundary values. We define $\gamma(g_1, g_2; A) = (\delta\beta_0)(g_1, g_2; A) = -(\delta\omega)(g) + \gamma_0(g_1, g_2; A)$, and the extension \hat{G}^D of G^D by $C^\infty(\mathcal{A}_M, S^1)$, where \mathcal{A}_M is the space of \mathfrak{g} -valued 1-forms on M . Then the factor group $\hat{G}^M = \hat{G}^D / \mathcal{G}$ is well defined and $\hat{\mathfrak{g}}^M$ is its Lie algebra (cf. §4).

In §4 we discussed two representations of the group $\hat{G}^M \times \hat{G}^M$, $M = S^1$. They both have analogues in the higher-dimensional case, but the second one (6) does not even require any modification. The 1-cocycle $\beta(U, g_L, g_R)$ on $\mathcal{G} \times \mathcal{G}$, the Wess-Zumino effective action for $M = S^3$, contains 18 terms instead of 4 terms in the loop case (6). An important open problem is to describe the higher-dimensional analogue of the level l representation W_l of $\hat{G} \times \hat{G}$.

Again, in the case when $G = \text{SU}(n)$, the condition (6) defines the bundle over A/\mathcal{G} which arises naturally from the Dirac family of operators via the family index theorem [1]. Note however that our \mathcal{A} is the space of \mathfrak{g} -valued 1-forms on the manifold with boundary D^{d+1} , while Atiyah and Singer considered only the 1-forms on the manifold without boundary S^{d+1} . Another construction of the representation of \hat{G}^M for $G = \text{SU}(n)$ parallel to some extent to the spin representation of §5.2 has been given in [18].

We have seen that the parallel between the affine and higher-dimensional cases goes remarkably far. There is also an analogue of the diffeomorphism group and the corresponding cocycles are known as gravitational anomalies. To find an explicit realization of the irreducible level 1 representations of extended current algebras, presumably analogous to the vertex operator construction, is an important open problem. In particular, after the affine case, the four-dimensional case can be of great importance in quantum field theory.

10. Conclusion. In this talk I wanted first and foremost to show the canonical origins and structures of affine Lie algebras, groups, and their representations and ways of generalizing them. All the motivations, analogies, and methods were dictated by the internal development of this theory parallel to a great extent to the classical finite-dimensional theory. The fact that this rather narrow approach leads to so many different mathematical structures indicates its enormous unifying power, which is not yet fully explored. Even more mysterious is that this purely mathematical point of view leads us into the very core of theoretical physics. In fact, the representation theory of affine Lie algebras and the

closely related Virasoro algebra is indistinguishable from the significant part of the correctly understood two-dimensional quantum field theory, the starting point of theoretical particle physics. Higher-dimensional generalizations of affine Lie algebras amount to the same structures as chiral and gravitational anomalies. Algebras of vertex operators and cohomology of the Virasoro algebra lead to a gauge invariant interacting string theory. Besides, discrete series of unitary representations of the Virasoro algebra uniformly explain critical phenomena in statistical mechanical models. Chiral and gravitational anomalies in quantum field theory and string theory represent two major recent developments in theoretical physics. The fact that they both come from canonical generalizations of affine Lie algebras assures us that further progress in the representation theory of infinite-dimensional Lie algebras and groups can be of invaluable importance in theoretical physics. We return to the old, and repeatedly arising, question about "the unreasonable effectiveness of mathematics in the natural sciences" (E. Wigner) and to the unavoidable conclusion that if theoretical physics does reflect the physical world, then the latter is the most canonical mathematical object.

Acknowledgements. I dedicate this talk to my friends in Leningrad who were the first critics of the points of view presented here in its early stages. I gratefully acknowledge Howard Garland and Gregg Zuckerman for their help and support through numerous discussions.

This research was done with the support of the National Science Foundation Grant #DMS-8602091 and the Sloan Foundation Grant #BR-2471T.

REFERENCES

As a source of references we use "A Kac-Moody Bibliography and Some Related References," compiled by G. Benkart and published in the Canadian Mathematical Society Conference Proceedings, volume 5, *Lie Algebras and Related Topics*, D. Britten, F. Lemire, and R. Moody, editors; in the text, we have indicated the author's name and the letter of the appropriate reference. The references to the talks presented at the International Mathematical Congresses are marked by author's name and year of the Congress. The references listed below are marked as bold numbers.

1. M. F. Atiyah and I. M. Singer, *Dirac operators coupled to vector potentials*, Proc. Nat. Acad. Sci. U.S.A. **81** (1984), 2597–2600.
2. K. Bardakci and M. B. Halpern, *New dual quark models*, Phys. Rev. D **3** (1971), 2493–2506.
3. R. E. Borcherds, *Vertex algebras, Kac-Moody algebras, and the Monster*, Proc. Nat. Acad. Sci. U.S.A. **83** (1986), 3068–3071.
4. R. C. Brower, *Spectrum-generating algebra and no-ghost theorem for the dual model*, Phys. Rev. D **6** (1972), 1655–1662.
5. M. Dubois-Viollette, M. Talon, and C. M. Viallet, *B.R.S. algebras. Analysis of the consistency equations in gauge theory*, Comm. Math. Phys. **102** (1985), 105–122.
6. L. D. Faddeev, *Operator anomaly for the Gauss law*, Phys. Lett. B **145** (1984), 81–84.
7. H. D. Fegan, *The heat equation on a compact Lie group*, Trans. Amer. Math. Soc. **246** (1978), 339–357.
8. I. B. Frenkel, H. Garland, and G. Zuckerman, *Semi-infinite cohomology and string theory*, Proc. Nat. Acad. Sci. U.S.A. **83** (1986), 8442–8446.

9. I. B. Frenkel, J. Lepowsky, and A. Meurman, *Vertex algebras and the Monster*, monography (in preparation).
10. I. M. Gelfand and D. B. Fuks, *The cohomology of the Lie algebra of the vector fields in a circle*, Funktsional. Anal. i Prilozhen **2** (1968), 92–93; English transl. in Functional Anal. Appl. **2** (1968), 342–343.
11. P. Goddard and C. B. Thorn, *Compatibility of the dual pomeron with unitarity and the absence of ghosts in the dual resonance model*, Phys. Lett. B **40** (1972), 235–238.
12. D. J. Gross, J. A. Harvey, E. Martinec, and R. Rohm, *Heterotic string theory*. I: *The free heterotic string*, Nucl. Phys. B **256** (1985), 253–284; II: *The interacting heterotic string*, Nucl. Phys. B **267** (1986), 75–124.
13. M. B. Halpern, *Quantum “solitons” which are $SU(N)$ fermions*, Phys. Rev. D **12** (1975), 1684–1699.
14. M. Kato and K. Ogawa, *Covariant quantization of string based on BRS invariance*, Nucl. Phys. B **212** (1983), 443–460.
15. O. Mathieu, *Classification des algebres de Lie graduées simples de croissance ≤ 1* , Invent. Math. **86** (1986), 371–426.
- 16a. J. Mickelsson, *Chiral anomalies in even and odd dimensions*, Comm. Math. Phys. **97** (1985), 361–370.
- 16b. —, *Kac-Moody groups, topology of the Dirac determinant bundle and fermionization*, Preprint.
17. J. L. Petersen, *Non-abelian chiral anomalies and Wess-Zumino effective actions*, Acta. Phys. Polon. B **16** (1985), 271–300.
18. G. Segal, *Faddeev’s anomaly in Gauss’s law*, Preprint.
19. I. M. Singer, *Families of Dirac operators with applications to physics*, *Elie Cartan et les Mathématiques d’aujourd’hui*, Astérisque, Special Issue, June 1986.
- 20a. E. Witten, *Non-abelian bosonization in two dimensions*, Comm. Math. Phys. **92** (1984), 455–472.
- 20b. —, *Non-commutative geometry and string field theory*, Nucl. Phys. B **268** (1986), 253–294.
21. B. Zumino, *Cohomology of gauge groups: cocycles and Schwinger terms*, Nucl. Phys. B **253** (1985), 477–493.

YALE UNIVERSITY, NEW HAVEN, CONNECTICUT 06520, USA

Geometrical Aspects of Representation Theory

VICTOR GINZBURG

I learned from I. M. Gel'fand that Mathematics of any kind is a Representation theory...

Yu. I. Manin

(from the short address on occasion of Professor Gel'fand's 70th birthday)

1. Geometry of primitive ideals.

1.1. Let G be a complex algebraic group with Lie algebra \mathfrak{g} . I would like to have a geometric picture for the structure of primitive ideals of the enveloping algebra $U(\mathfrak{g})$. The only universal answer known so far is the "orbit method," saying that primitive ideals should correspond to coadjoint orbits in the dual \mathfrak{g}^* of \mathfrak{g} . Sometimes this principle doesn't work, however. The situation may be improved if one looks not just onto an orbit itself but on a certain "resolution data" over it as well. I propose to consider diagrams like this:

$$\begin{array}{ccc} \mu^{-1}(G \cdot \lambda) & \hookrightarrow & T^\lambda X \\ \swarrow & & \swarrow \mu \quad \searrow \pi \\ G \cdot \lambda & \hookrightarrow & \mathfrak{g}^* \quad X = G/P \end{array} \quad (1.1.1)$$

Here $\lambda \in \mathfrak{g}^*$, P is a connected algebraic subgroup of G with Lie algebra \mathfrak{p} , such that $\lambda|_{[\mathfrak{p}, \mathfrak{p}]} = 0$, and $T^\lambda X := G \times_P (\lambda + \mathfrak{p}^\perp)$ is a "twisted cotangent bundle" on X (i.e., an affine bundle with symplectic structure on the total space, having Lagrangian fibres). The map μ is the "moment mapping," taking $(g, \xi) \in G \times (\lambda + \mathfrak{p}^\perp)$ to $(\text{Ad } g) \cdot \xi$. If \mathfrak{p} is a good polarization of λ so that the orbit method applies, then μ is an isomorphism of $T^\lambda X$ onto the orbit $G \cdot \lambda$ and $\pi \cdot \mu^{-1}(\lambda + \mathfrak{p}^\perp)$ is a single point of G/P .

DEFINITION 1.1.2. A solvable subgroup P , as above, is called a polarizing subgroup if the set $\pi \cdot \mu^{-1}(\lambda + \mathfrak{p}^\perp) \subset X$ consists of a finite number of P -orbits.

Let $F_x = G \cdot \lambda \cap \text{Ad } x \cdot (\lambda + \mathfrak{p}^\perp)$, $x \in X$. Given a *polarization* \mathfrak{p} (see [Di]), one gets a Lagrangian fibering: $G \cdot \lambda \rightarrow X$ with fibres F_x . There is no such fibering for a general polarizing subgroup P . Yet, one has

PROPOSITION 1.1.3. $\{F_x, x \in X\}$ is a G -invariant Lagrangian family in $G \cdot \lambda$. \square

Various F_x may intersect each other. However, after replacing $G \cdot \lambda$ by $\mu^{-1}(G \cdot \lambda)$, the family $\{F_x\}$ turns into the “resolved” one: $\tilde{F}_x := \mu^{-1}(G \cdot \lambda) \cap T_x^\lambda X$, induced by the standard family of fibres of the cotangent bundle.

PROPOSITION 1.1.4. $\mu^{-1}(G \cdot \lambda)$ is a coisotropic subvariety of $T^\lambda X$ and its projection onto $G \cdot \lambda$ coincides with the reduction map (relative to the 0-foliation on a coisotropic subvariety). \square

1.2. Given P , a polarizing subgroup, one can construct a finite number of \mathfrak{g} -modules. Set $F_0 = G \cdot \lambda \cap (\lambda + \mathfrak{p}^\perp)$. It follows from Propositions 1.1.3 and 1.1.4 that $\mu^{-1}(F_0)$ is a Lagrangian subvariety of $T^\lambda X$. Let Λ be one of its irreducible components and $\pi: \Lambda \rightarrow X$, the projection. Clearly, $\pi(\Lambda)$ is a P -stable irreducible subvariety of X . By Definition 1.1.2, there is a unique open orbit $P \cdot w \subset \pi(\Lambda)$, ($w \in X$), and Λ may be viewed as a “twisted conormal bundle” to $P \cdot w$. The space of distributions supported on $P \cdot w$ (with a certain twist depending on λ) is a $U(\mathfrak{g})$ -module, via the action of \mathfrak{g} on X . Let $L_{\lambda, \mathfrak{p}, \Lambda}$ be its $U(\mathfrak{g})$ -submodule generated by the zero-order P -eigendistributions, i.e., those having no derivatives in directions normal to the orbit.

THEOREM 1.2.1. Given $\lambda \in \mathfrak{g}^*$, one can find a polarizing subgroup P such that all $L_{\lambda, \mathfrak{p}, \Lambda}$ are nontrivial irreducible $U(\mathfrak{g})$ -modules (for all irreducible components Λ). \square

In particular, every $\lambda \in \mathfrak{g}^*$ has a polarizing subgroup.

1.3. From now on G is assumed to be a connected reductive group and \mathfrak{g}^* is identified with \mathfrak{g} via the Killing form. Let B be a Borel subgroup of G , T a maximal torus of B , W the Weyl group of (G, T) , $X = G/B$ the flag manifold, \mathfrak{n} the nilpotent radical of $\text{Lie } B$, and $\rho = \text{half sum of the roots in } \mathfrak{g}/\text{Lie } B$.

We remark that B is a polarizing subgroup for all points of $\text{Lie } B$, since the number of B -orbits in G/B is finite. Thus, the machinery of §§1.1–1.2 applies. The case of a nilpotent orbit $G \cdot \lambda$ is most interesting. The moment map μ of (1.1.1) then turns into the Springer resolution:

$$\mu: \text{the usual cotangent bundle } T^*X \rightarrow N = \{n \in \mathfrak{g} \mid \text{ad } n \text{ is nilpotent}\} \quad (1.3.1)$$

Here is the translation of (1.1.2)–(1.2.1) into our present language:

$$\text{Each component of } G \cdot \lambda \cap \mathfrak{n} \text{ is a Lagrangian subvariety of } G \cdot \lambda; \quad (1.3.2)$$

$$\mu^{-1}(\mathfrak{n}) \text{ is the union of all conormal bundles } T_{X_w}^* X \text{ to the Bruhat cells } X_w = B \cdot w \cdot B/B \text{ (} w \in W \text{);} \quad (1.3.3)$$

$$\text{The closure of an irreducible component of } \mu^{-1}(G \cdot \lambda \cap \mathfrak{n}) \text{ equals } \overline{T_{X_w}^* X} \text{ for some } w \in W; \quad (1.3.4)$$

$$\text{If } \bar{\Lambda} = \overline{T_{X_w}^* X}, \text{ then } L_{\lambda, B, \Lambda} = L_w \text{ is the irreducible highest weight module with the highest weight } (-\rho - w \cdot \rho). \quad (1.3.5)$$

1.4. Given a module M over a filtered algebra, let $S(M)$ be its *characteristic variety*, that is, the support of the graded module $\text{gr } M$, associated with a good filtration on M . Let \mathcal{D}_X be the sheaf of holomorphic differential operators on

the flag manifold X , filtered by degree, as usual. The characteristic variety of a \mathcal{D}_X - (resp. $U(\mathfrak{g})$)-module is a subvariety of T^*X (resp. \mathfrak{g}^*). It is known, for instance, that $S(\mathcal{D}_X \otimes_{U(\mathfrak{g})} L_w)$ is a Lagrangian subvariety of T^*X and $S(L_w)$ is its image under the moment map μ [BB].

Let $I_w = \text{Ann } L_w$ be a primitive ideal of $U(\mathfrak{g})$. If $G = \text{SL}_n$, the particularly nice geometric classification of I_w 's in terms of characteristic varieties is expected:

$$S(\mathcal{D}_X \otimes L_w) = \overline{T_{X_w}^* X} \quad (\text{conjectural}); \quad (1.4.1)$$

$$S(\mathcal{D}_X / \mathcal{D}_X \cdot I) = G \cdot \overline{T_{X_w}^* X} \quad (\text{conjectural}); \quad (1.4.2)$$

$$I_y \subset I_w \Leftrightarrow S(\mathcal{D}_X / \mathcal{D}_X \cdot I_y) \supset S(\mathcal{D}_X / \mathcal{D}_X \cdot I_w) \quad (\text{conjectural}); \quad (1.4.3)$$

$$S(U(\mathfrak{g})/I_w) = G \cdot (n \cap w \cdot n \cdot w^{-1}) \quad (\text{Joseph}). \quad (1.4.4)$$

All these statements but (1.4.3) are false whenever G is simple $\neq \text{SL}_n$. In general, one knows only that $S(U(\mathfrak{g})/I_w)$ is the closure of a nilpotent orbit [BB, KT, Gi2].

2. Representations of Weyl groups and Hecke algebras.

2.1. Why is it interesting to study the objects named in the title? The first reason is that irreducible representations of a Weyl group W have a beautiful geometric description, discovered by Springer [Spr], connecting them with nilpotent orbits in \mathfrak{g}^* (as in the “orbit method”!). Secondly, nilpotent orbits arise also as characteristic varieties of annihilators of simple \mathfrak{g} -modules. So, combining with Springer, one gets a strange relation: simple \mathfrak{g} -modules \leftrightarrow (simple) W -modules. That relation turns out to be crucial both in Joseph’s work on primitive ideals and in the work of Kazhdan-Lusztig [KL1]. Let me mention, as an illustration, that to finite-dimensional irreducible \mathfrak{g} -modules one attaches the “sign-representation,” appearing already in the Weyl character formula.

Another source of our interest originates from the observation that the Hecke algebra $H(q)$ may be viewed as a q -analogue of the group-algebra $\mathbf{Z}[W]$. Yet no Springer-type geometric theory for $H(q)$ -modules was known so far. Such a theory [Gi3] is given below for affine Hecke algebras together with a conjectural one for the ordinary Hecke algebra $H(q)$ (the “affine” case appears to be simpler). We begin with decomposing the two-sided regular representation of $H(q)$ or $\mathbf{Z}[W]$ into so-called “two-sided cell representations” (cf. §2.4). These cell representations of $H(q)$ are expected to coincide with those introduced by Kazhdan-Lusztig [KL1], while in the Weyl group case they provide an alternative simple approach to the Springer theory (cf. [Gi2]). Next, each two-sided cell representation corresponds to a nilpotent conjugacy class and irreducible constituents of such a representation form an “ L -packet,” parametrized by irreducible representations of a certain (often abelian) finite group. It should be emphasized that although the algebra $H(q)$ is (unnaturally) isomorphic to $\mathbf{C}[W]$ the cell representations for $H(q)$ and $\mathbf{C}[W]$ do not correspond to each other when $q \rightarrow 1$. Thus, there are two similar, but different, geometric pictures (for $q = 1$, resp. $q \neq 1$). I guess that the picture for $q \neq 1$, rather than Springer’s one, is relevant for the representation theory of complex Lie groups. Furthermore, it

is likely to be a correspondence: G -modules $\leftrightarrow H(q)$ -modules, similar to that known for reductive groups over finite fields (cf. §3) with $H(q)$ playing the role of an algebra of intertwining operators.

2.2. Keeping to the notations of §1.3, let $\tilde{N} = T^*X$ and $\Lambda = \tilde{N} \times_N \tilde{N} = \{(x_1, x_2) \in \tilde{N} \times \tilde{N} | \mu(x_1) = \mu(x_2)\}$. The group $\mathbf{C}^* \times G$ acts on N , \tilde{N} , and Λ (\mathbf{C}^* by multiplication and G by conjugation). Let $K_{\mathbf{C}^* \times G}(\Lambda)$ be the Grothendieck group of $\mathbf{C}^* \times G$ -equivariant coherent \mathcal{O}_Λ -sheaves.

We regard Λ as a correspondence from \tilde{N} to \tilde{N} . Clearly, $\Lambda \circ \Lambda = \Lambda$. There is a ring structure on $K_{\mathbf{C}^* \times G}(\Lambda)$ arising from composition of “sheaf-valued correspondences.” That is, given $[F], [F'] \in K_{\mathbf{C}^* \times G}(\Lambda)$, let $[F] \cdot [F']$ be the alternating sum of cohomology sheaves of the complex:

$$F \cdot F' = (p_{13})_*(p_{12}^*F \otimes_{\mathcal{O}_{\tilde{N} \times \tilde{N} \times \tilde{N}}} p_{23}^*F'),$$

where $p_{ij}: \tilde{N} \times \tilde{N} \times \tilde{N} \rightarrow \tilde{N} \times \tilde{N}$ is the projection along the factor not named. Thus, $K_{\mathbf{C}^* \times G}(\Lambda)$ becomes a $\mathbf{Z}[q, q^{-1}]$ -algebra with $\mathbf{Z}[q, q^{-1}]$ -action arising from that of the representation ring $R(\mathbf{C}^*) = \mathbf{Z}[q, q^{-1}]$.

For a maximal torus T of G let $\text{Hom}(T, \mathbf{C}^*)$ be the lattice of weights and $\tilde{W} := W \ltimes \text{Hom}(T, \mathbf{C}^*)$, the semidirect product (the “affine Weyl group”). Let H , resp. \tilde{H} , be the Hecke algebras attached to W , resp. \tilde{W} (cf. [KL1]).

THEOREM 2.2.1 [Gi3]. *There is an algebra isomorphism $\tilde{H} \simeq K_{\mathbf{C}^* \times G}(\Lambda)$. \square*

When viewed as a subvariety of $\tilde{N} \times \tilde{N} \cong T^*(X \times X)$, Λ becomes the union of conormal bundles to all G -orbits in $X \times X$. Hence, Λ has $\#W$ irreducible components, all of the same dimension $2d = 2 \cdot \dim X$. The group $CH_{2d}(\Lambda)$ of $2d$ -dimensional cycles in Λ is generated by these components. It also has a ring structure, so that the assignment $F \mapsto$ “ $2d$ -dimensional part of $\text{supp } F$ ” gives rise to a ring homomorphism $\text{supp}: K_G(\Lambda) \rightarrow CH_{2d}(\Lambda)$. Theorem 2.2.1 yields the following commutative diagram of ring homomorphisms:

$$\begin{array}{ccccc}
 H & \xrightarrow{i} & \tilde{H} & \xrightarrow[\text{(2.2.1)}]{\sim} & K_{\mathbf{C}^* \times G}(\Lambda) \\
 \downarrow \scriptstyle q \atop \downarrow \scriptstyle 1 & & \downarrow \scriptstyle q \atop \downarrow \scriptstyle 1 & & \downarrow \scriptstyle \text{forgetting} \\
 \mathbf{Z}[W] & \xrightarrow{i} & \mathbf{Z}[\tilde{W}] & \xrightarrow{\sim} & K_G(\Lambda) \\
 \text{id} \searrow & & \downarrow \scriptstyle r & & \downarrow \scriptstyle \text{supp} \\
 & & \mathbf{Z}[W] & \xrightarrow{\sim} & CH_{2d}(\Lambda)
 \end{array} \tag{2.2.2}$$

where the embeddings i are induced by the natural inclusion $W \hookrightarrow \tilde{W}$ and the projection r is induced by the homomorphism $\tilde{W} \rightarrow W$, taking $\text{Hom}(T, \mathbf{C}^*)$ to 1.

2.3. Consider the projection $\mu_\Delta: \Lambda = \tilde{N} \times_N \tilde{N} \rightarrow N$. For a G -stable closed subvariety $Y \subset N$ let $\tilde{H}(Y)$ (resp. $\mathbf{Z}[\tilde{W}](Y)$ or $\mathbf{Z}[W](Y)$) be the two-sided ideal of $\tilde{H} = K_{\mathbf{C}^* \times G}(\Lambda)$ (resp. $\mathbf{Z}[\tilde{W}]$ or $\mathbf{Z}[W]$) generated by sheaves or cycles supported on $\mu_\Delta^{-1}(Y)$. Given a nilpotent conjugacy class $C \subset N$, set: $\tilde{H}(C) = \tilde{H}(\overline{C})/\tilde{H}(\overline{C} - C)$, $\mathbf{Z}[\tilde{W}](C) = \mathbf{Z}[\tilde{W}](\overline{C})/\mathbf{Z}[\tilde{W}](\overline{C} - C)$, etc. These

spaces are bimodules over the algebras in question. The following conjecture is very close to that of [Lu3].

CONJECTURE 2.3.1. The $\tilde{H}(C)$'s are precisely the two-sided cell representations of \tilde{H} in the sense of [KL1].

Next, identify H with a subalgebra of \tilde{H} and set

$$H(C) = \tilde{H}(\overline{C}) \cap H / \tilde{H}(\overline{C} - C) \cap H.$$

It is likely that these are precisely the two-sided cell representations of H .

CONJECTURE 2.3.2. $H(C) \neq 0 \Leftrightarrow C$ is a "special" orbit (in the sense of Lusztig).

If true, this would provide a geometric understanding of "special" orbits. I also think that the reason for the characteristic variety $S(\mathcal{D}_X \otimes L_w)$ to be greater than $\overline{T_{X_w}^* X}$ is this: it might be no $\mathbf{C}^* \times G$ -equivariant sheaves contained in $H \subset \tilde{H}$ and supported on a single conormal bundle at the same time.

2.4. The center of $\mathbf{C} \otimes \tilde{H}$ is known to be isomorphic to $\mathbf{C}[T]^W[q, q^{-1}]$, where $\mathbf{C}[T]$ denotes the regular ring of T , a maximal torus. Let $I_{t,h}$ be the maximal ideal of the center, corresponding to a point $m = (t, h) \in \mathbf{C}^* \times T$ and let $\tilde{H}(t, h) := \mathbf{C} \otimes \tilde{H} / I_{t,h} \cdot \tilde{H}$ be the specialization of \tilde{H} at $m = (t, h)$.

Let $\mu: \tilde{N}^m \rightarrow N^m$ be the restriction of the moment map: $\tilde{N} \rightarrow N$ to m -fixed point subvarieties. For $n \in N^m$, the fibre $X_n^h := \mu^{-1}(n) \subset \tilde{N}^m$ may be viewed as a variety of all Borel subgroups of G , containing both h and $\exp n$. The group $Z_G(h, n)$ acts on it by conjugation, giving rise to an action of the finite group $A(h, n) = Z_G(h, n) / Z_G^0(h, n)$ on $H_*(X_n^h)$.

PROPOSITION 2.4.1 [Gi3, KL3]. *There is an $\tilde{H}(t, h)$ -action on $H_*(X_n^h)$, commuting with that of $A(h, n)$. \square*

This $\tilde{H}(t, h)$ -action is a q -analogue of a Weyl group action, defined by Springer [Spr].

PROPOSITION 2.4.2 [Gi3]. *There exists an $A(h, n)$ -invariant symmetric form (\cdot, \cdot) on $H_*(X_n^h)$, which is "contravariant" with respect to the $\tilde{H}(t, h)$ -action; i.e., $(a \cdot x, y) = (x, a^* \cdot y)$ for $x, y \in H_*(X_n^h)$ and an anti-involution: $a \mapsto a^*$ on \tilde{H} , taking T_w to $T_{w^{-1}}$ when $w \in W \subset \tilde{W}$.*

The form (\cdot, \cdot) is defined by intersecting cycles $x, y \in H_*(X_n^h)$ in an ambient (smooth) variety S , the inverse image of a transversal slice to an orbit in N^m .

It follows from Propositions 2.4.1 and 2.4.2 that there is an orthogonal decomposition:

$$H_*(X_n^h) = \bigoplus_{\chi} K_{h,n,\chi} \otimes \chi, \quad (2.4.3)$$

where χ runs over irreducible representations of $A(h, n)$ and $K_{h,n,\chi}$ are certain $\tilde{H}(t, h)$ -modules.

2.5. The representation theory of $\tilde{H}(t, h)$ that I am going to present amazingly resembles that of highest weight \mathfrak{g} -modules (i.e., category \mathcal{O} of [BGG]). In both cases, three types of objects play the prominent role: simple modules L_α ,

their projective covers P_α , and the intermediate “standard” modules $K_{h,n,\chi}$ (cf. (2.4.3)), which are the counterparts of the Verma modules. The \tilde{H} -module theory has an advantage of giving way to direct geometric interpretation. The module $K_{h,n,\chi}$, for instance, lives in homology of X_n^h and has a contravariant form (\cdot, \cdot) , arising from intersecting cycles (cf. Proposition 2.4.2). Just as in the \mathfrak{g} -module case, all simple \tilde{H} -modules turn out to be of the form: $L_\alpha = K_{h,n,\chi}/\text{rad}(\cdot, \cdot)$, for some triple (h, n, χ) , which is unique up to conjugation by G . So we write: $\alpha = (h, n, \chi)$. The multiplicities $(K_\alpha : L_\beta)$ and $(P_\alpha : L_\beta)$ are in both cases related to each other by similar formulas (cf. Proposition 2.5.4 and [BGG]), and can be expressed in terms of the intersection cohomology (Kazhdan-Lusztig type formula; cf. §§2.5.4–2.6.2 and [BK, BeBe]). Furthermore, the derived category of graded $\tilde{H}(t, h)$ -modules is equivalent to the derived category of *mixed* complexes on N^m (cf. Theorem 2.6.3 and [BG]) so that grading on modules corresponds to the weight filtration on sheaves. This is an \tilde{H} -counterpart of the well-known equivalence between \mathfrak{g} -modules and perverse sheaves (on the flag manifold), established via \mathcal{D} -modules [BeBe, BK].

Fix $m = (t, h)$, a semisimple element of $\mathbf{C}^* \times G$. Let

$M(m) = \{Z_G(h)\text{-conjugacy classes of } \alpha = (h, n, \chi): (\text{Ad } h) \cdot n = t^{-1} \cdot n, \text{ with } \chi \text{ an irreducible representation of } A(h, n) \text{ occurring in } H_*(X_n^h)\}.$

For $\alpha = (h, n, \chi) \in M(m)$ the group $A(h, n)$ may be viewed as a quotient of $\pi_1(O_\alpha)$, where $O_\alpha = Z_G(h) \cdot n$ is an orbit in N^m . Let \mathcal{L}_α be the intersection cohomology complex on \overline{O}_α with the monodromy χ . By the *decomposition theorem* [BBD], applied to the projective map $\mu: \tilde{N}^m \rightarrow N^m$, we get (cf. [BM]):

$$\mathbf{R}\mu_*(\mathbf{C}_{\tilde{N}^m}) = \bigoplus_{\alpha \in M(m)} L_\alpha \otimes \mathcal{L}_\alpha \quad (\text{shifts disregarded}), \quad (2.5.1)$$

where L_α are certain vector spaces. It turns out that: $L_\alpha \cong K_\alpha/\text{rad}(\cdot, \cdot)$.

Let $H_*(\Lambda^m, \mathbf{C})$ be the Borel-Moore homology of the m -fixed point subvariety $\Lambda^m \subset \Lambda$. Theorem 2.2.1 and (2.5.1) yield

PROPOSITION 2.5.2 [Gi3]. *There are algebra isomorphisms:*

$$\tilde{H}(t, h) \cong H_*(\Lambda^m, \mathbf{C}) \cong \bigoplus_{i \geq 0, \alpha, \beta} \text{Hom}(L_\alpha, L_\beta) \otimes \text{Ext}^i(\mathcal{L}_\alpha, \mathcal{L}_\beta). \quad \square$$

THEOREM 2.5.3 [Gi3] (cf. [KL3]). *Suppose t is not a root of unity. Then $\{L_\alpha | \alpha \in M(m), m = (t, h)\}$ is precisely the collection of all simple $\tilde{H}(t)$ -modules.* \square

Given two orbits $O_\beta \subset \overline{O}_\alpha$, $\beta = (h, n, \chi)$, consider the χ -component of the stalk $(\mathcal{H}^i \mathcal{L}_\alpha)_n$, and set $\mathcal{L}_{\alpha, \beta}^i = \dim(\mathcal{H}^i \mathcal{L}_\alpha)_n^\chi$.

Let \mathcal{L} be the unipotent matrix with entries $\mathcal{L}_{\alpha, \beta} = \sum_i \mathcal{L}_{\alpha, \beta}^i$, D the diagonal matrix: $D_{\alpha, \alpha} = \text{Euler characteristic of } O_\alpha$ (with compact support), and $(K : L)$, $(P : L)$ the multiplicity matrices with entries $(K_\alpha : L_\beta)$, resp. $(P_\alpha : L_\beta)$.

PROPOSITION 2.5.4. $(K : L) = \mathcal{L}$ and $(P : L) = \mathcal{L} \cdot D \cdot \mathcal{L}^t$. \square

2.6. Following an idea of Jantzen, one deforms the parameter $(h, n, \chi) = \alpha$ of a standard module in order to get a module \tilde{K}_α over $\mathbb{C}[u]$, the polynomial ring, so that $K_\alpha = \tilde{K}_\alpha/u \cdot \tilde{K}_\alpha$. The contravariant form on K_α lifts to a $\mathbb{C}[u]$ -valued form on \tilde{K}_α . The Jantzen filtration J^\bullet on standard modules is defined by:

$$J^k K_\alpha = \{x \in \tilde{K}_\alpha \mid (x, \tilde{K}_\alpha) \subset u^k \cdot \tilde{K}_\alpha\} / (u). \tag{2.6.1}$$

Again, the construction has a direct geometric interpretation via \mathbb{C}^* -equivariant cohomology (recall that $H^*(B_{\mathbb{C}^*}) \cong H^*(\mathbb{CP}^\infty) \cong \mathbb{C}[u]$). The Jantzen filtration then turns out to be related to intersecting with a “hyperplane section.” The Jantzen filtration on Verma modules is related, in the same way, to the monodromy on vanishing cycles, thus supporting the analogy [De] between the monodromy and the Lefschetz operators.

THEOREM 2.6.2 [Gi3]. $\mathrm{Gr}_J K_\alpha$ (see (2.6.1)) is a semisimple $\tilde{H}(t, h)$ -module and $(\mathrm{Gr}_J^i K_\alpha : L_\beta) = \mathcal{L}_{\alpha, \beta}^{-i}$, $i = 0, 1, \dots$. \square

Consider the grading on $\tilde{H}(t, h)$ given by the index “ i ” in the direct sum decomposition in Proposition 2.5.2. Using the assignment $\mathcal{L}_\alpha \rightsquigarrow P_\alpha$ one obtains:

THEOREM 2.6.3 [BG]. The category of mixed perverse sheaves on N^m , having their Jordan-Hölder factors among \mathcal{L}_α , ($\alpha \in M(m)$), is equivalent to the category of bounded complexes:

$$\cdots \rightarrow P_i \rightarrow P_{i+1} \rightarrow \cdots,$$

where P_i denotes a graded projective $\tilde{H}(t, h)$ -module generated by elements of degree i .

2.7. Given $\alpha = (h, n, \chi) \in M(m)$ let $C = G \cdot n$. The action of \tilde{H} on L_α gives rise to an algebra homomorphism (notation of §2.3): $\tilde{H}(\overline{C}) \rightarrow \mathrm{End}(L_\alpha)$, vanishing on $\tilde{H}(\overline{C}-C)$. Let $f: H(C) \hookrightarrow \tilde{H}(C) \rightarrow \mathrm{End}(L_\alpha)$ be the induced homomorphism. The algebra $H(n, h, \chi) := \mathbb{C} \otimes f(H(C))$ doesn’t change when $(t, h) \in \mathbb{C}^* \times G$ varies in a connected component of the isotropy subgroup of n , for it is semisimple, hence rigid. Thus, $H(n, h, \chi)$ depends only on the conjugacy class of (n, \bar{h}, χ) , where \bar{h} is the image of (t, h) in $\pi_1(C)$ and χ is viewed as an irreducible representation of its centralizer in $\pi_1(C)$.

CONJECTURE 2.7.1. $H(n, h, \chi)$ is a simple algebra.

Moreover, the assignment $(n, \bar{h}, \chi) \mapsto H(n, h, \chi)$ (for n special, cf. Conjecture 2.3.2) seems to provide a parametrization of simple H -modules.

3. Towards the “Langlands geometry.” Throughout this section G stands for a connected split reductive group with connected center and G^* for the dual group in the sense of [La].

3.1. Let \mathbb{F}_q be a finite field. The classification of irreducible complex representations of the finite group $G(\mathbb{F}_q)$ [Lu2] splits into two parts. Given a simple $G(\mathbb{F}_q)$ -module, one attaches to it a semisimple conjugacy class in $G^*(\mathbb{F}_q)$ and a “unipotent” representation of a smaller reductive group [Lu2]. The classification of unipotent representations was carried out in [Lu2]. As presented there,

it appears to me quite mysterious. A part of it, at least, has a simple geometric explanation. Namely, consider irreducible representations arising from decomposition of the “principal series representation” $V = \mathbf{C}[G(\mathbf{F}_q)/B(\mathbf{F}_q)]$, where $B(\mathbf{F}_q)$ is a split Borel subgroup. The ring $\text{Hom}_{G(\mathbf{F}_q)}(V, V)$ is spanned by the double-cosets relative to $B(\mathbf{F}_q)$ and is isomorphic to the Hecke algebra $\mathbf{C} \otimes H(q)$. Thus, the irreducible constituents of V are in (1-1)-correspondence with simple $H(q)$ -modules and the latter were (conjecturally) classified in §2.7.

3.2. Let K be a p -adic field with the residue class field \mathbf{F}_q . By the general “Langlands philosophy” [La], irreducible representations of $G(K)$ should correspond to conjugacy classes of homomorphisms: $\text{Gal}(\overline{K}/K) \rightarrow G^*(\mathbf{C})$, where $\text{Gal}(\overline{K}/K)$ denotes the Galois-Weil group. Following [Lu1], one considers the refined data $\alpha = (C, \chi)$, consisting of such a conjugacy class C and an irreducible character χ of $\pi_1(C)$.

Let Γ be the quotient of $\text{Gal}(\overline{K}/K)$, corresponding to the maximal tamely ramified extension of K . The group Γ has the generators F (Frobenius) and M (Monodromy), subject to the relation: $F \cdot M \cdot F^{-1} = M^q$. Hence, the refined data attached to a homomorphism $\Gamma \rightarrow G^*(\mathbf{C})$ is determined by a triple $\alpha = (h, g, \chi)$, such that $h \cdot g \cdot h^{-1} = g^q$ ($h, g \in G^*(\mathbf{C})$). Deligne and Langlands conjectured (cf. [Lu1]) that:

Irreducible representations of $G(K)$ occurring in the “unramified principal series” are in (1-1)-correspondence with the conjugacy classes of triples $\alpha = (h, g, \chi)$ with h semisimple and g unipotent. (3.2.1)

As in §3.1, the irreducible representations in question are in (1-1)-correspondence with simple modules over $\tilde{H}(q)$, the “affine” Hecke algebra, attached to $G^*(\mathbf{C})$. Thus, the conjecture (3.2.1) follows from Theorem 2.5.3 (with g replaced by $n = \log g$).

I believe that a similar mechanism is responsible for the general Langlands conjecture. Namely, there should be a variety \tilde{G}^* , a pure perverse sheaf \mathcal{C} on \tilde{G}^* and a proper map $\mu: \tilde{G}^* \rightarrow \text{Hom}(\text{Gal}(\overline{K}/K), G^*(\mathbf{C}))$ so that irreducible representations of $G(K)$ correspond to the irreducible constituents of $\mathbf{R}\mu_* \mathcal{C}$ (cf. (2.5.1)). Furthermore, an appropriate derived category of admissible $G(K)$ -modules is expected to be equivalent to a derived category of mixed constructible complexes on $\text{Hom}(\text{Gal}(\overline{K}/K), G^*(\mathbf{C}))$.

REFERENCES

- [BeBe] A. Beilinson and J. Bernstein, *Localization de g -modules*, C. R. Acad. Sci. Paris **292** (1981), 15–18.
- [BBD] A. Beilinson, J. Bernstein, and P. Deligne, *Faisceaux pervers*, Asterisque, no. 100, Soc. Math. France, Paris, 1982.
- [BG] A. Beilinson and V. Ginsburg, *Mixed sheaves and Hecke algebras*, Preprint, Moscow, 1987.
- [BG2] ———, *Mixed categories, Ext-duality and representations*, Preprint, Moscow, 1986.

- [BGG] I. N. Bernstein, I. M. Gel'fand, and S. I. Gel'fand, *A category of g -modules*, Functional Anal. Appl. **10** (1976), 1–18.
- [BB] W. Borho and J.-L. Brylinski, *Differential operators on homogeneous spaces*. I, II, Invent. Math. **69** (1982), 437–476; **80** (1985), 1–68.
- [BM] W. Borho and R. MacPherson, *Representations des groupes de Weyl et homologie d'intersection pour les varietes nilpotentes*, C. R. Acad. Sci. Paris **292** (1981), 707–710.
- [BK] J.-L. Brylinski and M. Kashiwara, *Kazhdan-Lusztig conjecture and holonomic systems*, Invent. Math. **64** (1981), 387–410.
- [De] P. Deligne, *Théorie de Hodge*. I, Proc. Internat. Congr. Math. (Nice, 1970), Vol. 1, Gauthier-Villars, Paris, 1971, pp. 425–430.
- [Di] J. Dixmier, *Algèbres enveloppantes*, Gauthier-Villars, Paris, 1974.
- [G1] V. Ginsburg, *Characteristic varieties and vanishing cycles*, Invent. Math. **84** (1986), 327–402.
- [G12] —, *g -modules, Springer's representations and bivariant Chern classes*, Adv. in Math. **61** (1986), 1–48.
- [G13] —, *Deligne-Langlands conjecture and representations of affine Hecke algebras*, Preprint, Moscow, 1985.
- [KL1] D. Kazhdan and G. Lusztig, *Representations of Coxeter groups and Hecke algebras*, Invent. Math. **53** (1979), 165–184.
- [KL2] —, *A topological approach to Springer's representations*, Adv. in Math. **38** (1980), 222–228.
- [KL3] —, *Proof of the Deligne-Langlands conjecture for Hecke algebras*, Invent. Math. **87** (1987), 153–215.
- [La] R. Langlands, *Problems in the theory of automorphic forms*, Lecture Notes in Math., vol. 170, Springer-Verlag, 1970, pp. 18–86.
- [Lu1] G. Lusztig, *Some examples of square integrable representations of p -adic groups*, Trans. Amer. Math. Soc. **277** (1983), 623–653.
- [Lu2] —, *Characters of reductive groups over a finite field*, Ann. of Math. Studies, vol. 107, Princeton Univ. Press, Princeton, N.J., 1984.
- [Lu3] —, *Cells in affine Weyl groups*, Algebraic Groups and Related Topics, Adv. Stud. Pure Math., vol. 6, Kinokunya, Tokyo and North-Holland, Amsterdam, 1985.
- [Mac] R. MacPherson, *Global questions in the topology of singular spaces*, Proc. Internat. Congr. Math. (Warsaw, 1983), North-Holland, 1984, pp. 213–235.
- [Spr] T. A. Springer, *Trigonometric sums, Green functions of finite groups and representations of Weyl groups*, Invent. Math. **36** (1976), 173–207.

MOSCOW STATE UNIVERSITY, 117234 MOSCOW, USSR

Algebraic Geometry and Representation Theory

DAVID KAZHDAN

0. Let \mathbf{G} be a reductive group over \mathbf{Z} . For any field F we can consider the group G_F of F -points on \mathbf{G} . At first glance, the groups G_F for different fields F appear to have little in common with each other. I. Gelfand has conjectured that

(1) The structure of the representations of G_F has fundamental features which do not depend on a choice of F .

(2) Moreover, it is possible to define representations by formulas which are universally valid over any local or finite fields.

In the book [GGP-S] both conjectures are proved for $G = SL_2$ (see Chapter 2, §§4.1 and 5.4). Unfortunately the second conjecture is not known for any other group.

Langlands reformulated the first conjecture in a more precise form [B], and it has been proven in a number of cases.

Since almost nothing is known about the second conjecture, I will postpone its discussion until the very end of this paper and will restrict myself to applications of Algebraic Geometry to Representation Theory.

1. Finite fields. Let F be a finite field of characteristic p , \mathbf{G} a reductive F -group, and $G = G_F$. Let $B = T_0U \subset G$ be a Borel subgroup. For any character χ of T_0 , we denote by π_χ the induced representation $\pi_\chi = \text{Ind}_{B_0}^G \chi$. Such representations are called "the principal series representations." We say that a character χ is generic if $\chi^w \neq \chi$ for any nontrivial element w of the Weyl group W of G . It is easy to see that

(1) for any generic character $\chi: T_0 \rightarrow \mathbb{C}^*$, the induced representation π_χ is irreducible;

(2) the representations $\pi_\chi, \pi_{\chi'}$ are equivalent if and only if χ, χ' are conjugate under the action of the Weyl group W ;

(3) for any generic $t_0 \in T_0$ we have $\text{Tr } \pi_\chi(t_0) = \sum_{w \in W} \chi(t_0^w)$.

Macdonald has conjectured that for any Cartan subgroup $T \subset G$ and any generic character $\chi: T \rightarrow \mathbb{C}^*$, there exists a unique irreducible representation π_χ of G , such that for any generic $t \in T$ we have $\text{Tr } \pi_\chi(t) = \pm \sum_{w \in W_T} \chi(t^w)$, where

$W_T = \text{Norm}_G(T)/T$. Springer [Sp1] suggested a formula for the character of $\pi_\chi(g)$ for arbitrary g in G . To formulate his conjecture we assume that p is sufficiently large. Let \mathfrak{G} be the Lie algebra of G , $\Omega \subset \mathfrak{G}$ a regular conjugacy class containing an element of the Lie algebra of T , and ϕ_Ω the Fourier transform of the characteristic function of Ω . Springer [Sp2] proved that the restriction of ϕ_Ω to the set of nilpotent elements in \mathfrak{G} does not depend on the choice of Ω ; he conjectured that $\text{Tr } \pi_\chi(u) = \phi_\Omega(\log u)$ for any unipotent u in G . This conjecture was proved a little later [K]. Both Springer's proof of Ω -independence and the proof of his conjecture are based on the algebro-geometric reduction to the case of principle series, where $T = T^0$.

At first glance, the formula for $\text{Tr } \pi_\chi$ is given piecemeal—one formula for unipotent elements, a completely different formula for semisimple elements. But Lusztig realized the existence of a natural algebro-geometric object describing $\text{Tr } \pi_\chi$. He showed that there exists a unique irreducible perverse sheaf \mathfrak{F}_χ over G such that for any g in G we have that $\text{Tr } \pi_\chi(g)$ equal to the trace of the Frobenius at g on the fiber of \mathfrak{F}_χ at g . Later he showed that the characters of all irreducible representations of G can be expressed in terms of perverse sheaves.

Another application of Algebraic Geometry to representation theory of finite Lie groups is the realization of representations of G in cohomology groups of algebraic varieties. It was discovered by Drinfeld (in the case $G = GL_n$) and by Deligne and Lusztig [D-L] (in the general case). This discovery led Lusztig ultimately to the classification of irreducible representations of reductive groups over a finite field [L].

2. Local fields. In the case when $F = \mathbb{R}$ or \mathbb{C} , the theory of D -modules [V] gave the possibility of finding characters of all irreducible representations of G . In the case when F is a nonarchimedean field, K -theory provides the way to construct all the irreducible components of the unramified principle series [K-L]. This topic will be discussed in more details in V. Ginzburg's talk (see his preprint [G]).

Another application of Algebraic Geometry to representation theory of local groups is the work of Henniart (see his letter to Laumon) in which he proves the (weak) form of the local Langlands conjecture for GL_n . The proof is based on Laumon's theory of the Fourier transform of representations of Galois groups of local fields with positive characteristic.

3. Global fields. Let F be a global field of finite characteristic p . In this case, F is the field of rational functions on a curve C over a finite field \mathbb{F}_q . Let \mathbf{A} be the ring of adèles of F and let $O \subset \mathbf{A}$ be the subring of integral adèles. A special case of Langlands' conjecture states the existence of a natural one-to-one correspondence $\rho \rightarrow \pi_\rho$ between the set of equivalence classes of irreducible n -dimensional l -adic representations of $\pi_1(C)$, $l \neq p$, and the set of equivalence classes of unramified cuspidal automorphic representations of $GL_n(\mathbf{A})$. This correspondence was established by V. Drinfeld for $n = 2$ [D] in the following way.

Let $\rho: \pi_1(C) \rightarrow GL_2(Q_l)$ be an irreducible representation. Drinfeld associates to ρ a perverse sheaf \mathfrak{F}_ρ on the space \mathfrak{M}_2 of 2-dimensional vector bundles over C . The trace of the Frobenius on \mathfrak{F}_ρ defines a function ϕ_ρ on the set $\mathfrak{M}_2(F_q)$. Since $\mathfrak{M}_2(F_q) \simeq GL_2(O) \backslash GL_2(\mathbf{A}) / GL_2(F)$ (see [W]) we may consider ϕ_ρ as a function on $GL_2(\mathbf{A}) / GL_2(F)$. Drinfeld showed that ϕ_ρ is an automorphic forms which corresponds to π_ρ .

Another application of Algebraic Geometry to automorphic forms in positive characteristic is the proof of the Ramanujan conjecture for automorphic representations of $GL_n(\mathbf{A})$ with at least one supercuspidal component.

4. Different fields. Until now we have discussed different types of fields separately. But they are very much related. For example, the notion of a unipotent representation for groups over finite fields [D-L] has a very precise analogue for representations of real groups [BV].

5. A possibility. We were describing applications of Algebraic Geometry to Representation Theory. Now we try to envisage the possible algebro-geometric structure of Representation Theory. We will restrict ourselves to the case where F is a local field. Let \underline{G} be a reductive group over \mathbf{Z} .

It is difficult to imagine the existence of “an object” \hat{G} over \mathbf{Z} such that irreducible representations of G_F correspond to “ F -points” of \hat{G} for local fields F . Even in the simplest case $\mathbf{G} = \mathbf{G}_m$, the corresponding functor $F \rightarrow \{\text{Multiplicative characters of } F\}$ does not have an obvious algebro-geometric interpretation. On the other hand, we can try to describe “series of representations.” Let \mathbf{H} , \mathbf{G} be quasisplit reductive \mathbf{Z} -groups and let $\phi: {}^L\mathbf{H} \rightarrow {}^L\mathbf{G}$ be a morphism of the corresponding L -groups (see [B]). Langlands’ philosophy predicts the existence of a correspondence between representations of H_F and G_F for any (local) F . We can try to imagine a “materialization” of this correspondence in the following way.

Does there exist a \mathbf{Z} -variety X , a differential form ω of highest degree on X , and rational functions R^ϕ , K^ϕ on $H \times G \times X \times X$ such that for any local field F and a nontrivial additive character ψ on F , the operators $T(h \times g)$ on $L^2(X_F, |\omega|_F)$, given by the formula

$$T(h \times g)(f)(x) \stackrel{\text{def}}{=} \int_{y \in X_F} \left(\frac{\psi(K^\phi(h, g, x, y))}{|R^\phi(h, g, x, y)|} \right) f(y) |\omega|_F,$$

define a projective unitary representation of $H_F \times G_F$ which realizes the Langlands correspondence? An example of such a construction is in [GGP-S]. In this case $\mathbf{G} = SL_2$, and \mathbf{H} is the group of units in a quadratic extension E of \mathbf{Q} . We take $X = E$, ω to be a standard two form on X ($\approx \mathbf{A}^2$), $R \equiv 1$, and define K by the formula

$$K(h, g, x, y) = c^{-1}(aN(y) + dN(x) - \text{Tr } hxy)$$

for $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, where $N: E^* \rightarrow \mathbf{Q}^*$ and $\text{tr}: E \rightarrow \mathbf{Q}$ are the norm and the trace on E . It is natural to ask for which series of representations it is possible to realize Gelfand’s conjecture in such a form.

REFERENCES

- [B] A. Borel, *Automorphic L-functions*, Proc. Sympos. Pure Math., vol. 33, part 2, Amer. Math. Soc., Providence, R.I., 1979, pp. 27–61.
- [BV] D. Barbasch and D. Vogan, *Unipotent representations of complex semisimple groups*, Ann. of Math. **121** (1985), 41–110.
- [D] 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.
- [D-L] P. Deligne and G. Lusztig, *Representations of reductive groups over finite fields*, Ann. of Math. **103** (1976), 103–161.
- [G] V. Ginsburg, *Deligne-Langlands conjecture and representations of affine Hecke algebras*, Preprint.
- [GGP-S] I. Gelfand, M. Graev, and I. Pyatetskii-Shapiro, *Representation theory and automorphic forms*, Saunders, Philadelphia, Pa.-London-Toronto, Ont., 1969.
- [K] D. Kazhdan, *Proof of Springer's hypothesis*, Israel J. Math. **28** (1977), 272–286.
- [K-L] D. Kazhdan and G. Lusztig, *Proof of the Deligne-Langlands conjecture for Hecke algebras*, Invent. Math. **87** (1987), 153–215.
- [L] G. Lusztig, *Characters of reductive groups over a finite field*, Princeton Univ. Press, Princeton, N.J., 1984.
- [Sp1] T. Springer, *Generalization of Green polynomials*, Proc. Sympos. Pure Math., vol. 21, Amer. Math. Soc., Providence, R.I., 1971, pp. 149–153.
- [Sp2] ———, *Trigonometric sums*, Invent. Math. **36** (1976), 173–206.
- [V] D. Vogan, *Irreducible characters of semisimple Lie groups. III*, Invent. Math. **71** (1983), 381–417.
- [W] A. Weil, *Generalisation des fonctions abéliennes*, J. Math. Pures Appl. (9) **17** (1938), 47–87.

HARVARD UNIVERSITY, CAMBRIDGE, MASSACHUSETTS 02138, USA

On the Supercuspidal Representations of GL_N and Other p -adic Groups

P. C. KUTZKO

1. Introduction. Although I will be limiting myself here largely to the groups $GL_N(F)$ where F is a p -adic field, it will be useful to consider at first an arbitrary totally disconnected locally compact group G . To begin at the beginning, by a representation (π, V) of G I will mean a complex vector space V together with a homomorphism π of G into $\text{Aut}_{\mathbb{C}}(V)$. If (π, V) is such a representation, then a vector v in V is said to be *smooth* if its stability subgroup (via the action given by π) in G is open; the set V^∞ of smooth vectors in V is in fact a G -subspace of V and, in particular, (π, V) is said to be a *smooth representation* if $V = V^\infty$. (For generalities concerning smooth representations, see [Cs] or [Ct].)

Let \tilde{V} be the space of linear functionals on V . Then there is a natural representation, $\tilde{\pi}$ of G on \tilde{V} given by $(\tilde{\pi}(g)\varphi)(v) = \varphi(\pi(g^{-1})v)$ for v in V , φ in \tilde{V} . The representation $(\tilde{\pi}, \tilde{V})$ need not be smooth even if (π, V) is. In any event, one may consider the restriction, $\check{\pi}$, of $\tilde{\pi}$ to the space of smooth functionals $\check{V} = (\tilde{V})^\infty$; if (π, V) is itself smooth, then the representation $(\check{\pi}, \check{V})$ is called the *contragredient representation* of (π, V) .

If G were a finite group and (π, V) were a finite-dimensional representation of G , then a choice of basis in V would identify the endomorphism $\pi(g)$ with a matrix. One would, by considering the coefficients of this matrix as functions of G , obtain certain complex-valued functions on G called *matrix coefficients* of π . If, now, (π, V) is a smooth representation of a locally compact totally disconnected group G , then, by analogy with the above case, a *coefficient* of π is defined to be any function on G of the form $x \mapsto \varphi(\pi(x)v)$, where v lies in V and φ is a *smooth* functional on V . With this last definition, I may now describe the representations in which we will be interested in what follows.

(1.01) **DEFINITION.** A smooth representation (π, V) of a totally disconnected locally compact group G is said to be *supercuspidal* if each of its coefficients is supported on a set which is compact-modulo-center, that is, if given a coefficient c of π , there is a compact subset C of G so that the support of c is contained in the set $Z(G) \cdot C$, where $Z(G)$ is the center of G .

The main motivation for the study of supercuspidal representations is the role that they play in the study of the representation theory of p -adic groups. In this context we have the following fundamental results of Jacquet [J].

(1.02) Let G be a reductive group over a p -adic field F . Then a smooth irreducible representation (π, V) of G is supercuspidal if and only if for any proper parabolic subgroup P with unipotent radical N we have $\text{Hom}_N(\pi, 1_N) = 0$, where 1_N is the identity representation of N .

So, supercuspidal representations of p -adic groups are in direct analogy with cuspidal representations of finite linear groups (see, e.g., [HC]).

(1.03) Let π be a smooth irreducible representation of a reductive group G over a p -adic field F .

Then there is (not necessarily proper) parabolic subgroup $P = MN$ (Levi factorization) and a supercuspidal representation τ of M so that π is a subrepresentation of the representation $\text{Ind}_P^G(\tau \otimes 1_N)$ of G induced from $\tau \otimes 1_N$. (For a precise definition of induction in this context, see [S].)

The content of the above results is that the classification of the smooth irreducible representations of a reductive p -adic group G must in some sense begin with a classification of the irreducible supercuspidal representation of G . This latter classification can be approached in at least two ways, and it is the interplay between these approaches which I would like to explore here.

2. The Langlands correspondence in the case of $\text{GL}_N(F)$. I will not discuss the general Langlands correspondence here (see, e.g., [Bo]) but will, instead, take as my starting point the observation that the validity of this correspondence would imply that the set of (equivalence classes of) irreducible supercuspidal representations of the group $\text{GL}_N(F)$ would be naturally parametrized by the set of (equivalence classes of) irreducible N -dimensional representations of the absolute Weil group, $W(F)$, of F . Looking at this parametrization heuristically, that is, using facts about the structure of the set of N -dimensional irreducible representations of $W(F)$ to speculate about analogous properties of the set of irreducible supercuspidal representations of $\text{GL}_N(F)$, one is led to list several properties of the set of irreducible N -dimensional representations of $W(F)$:

(2.01) If E is any p -adic field, then the abelianization, $(W(E))_{\text{ab}}$, of $W(E)$ is canonically isomorphic with the multiplicative group, E^\times , of E . In particular, N -dimensional *monomial* representations of $W(F)$ may be thought of as being parametrized (maybe in several ways) by (not necessarily unitary) characters of E^\times , where E/F is an extension of degree N .

(2.02) If $p \nmid N$, then all N -dimensional representations of $W(F)$ are monomial (see, e.g., [D]).

(2.03) If θ is a character of E^\times and if E/F is galois, then for $\text{Ind}_{E/F} \theta$ (i.e., $\text{Ind}_{W(E)}^{W(F)} \theta$) to be irreducible, it is necessary and sufficient that $\theta^\tau \neq \theta$ for all nontrivial elements τ of $\text{Gal}(E/F)$. If, in particular, E/F is cyclic (i.e., $\text{Gal}(E/F)$ is cyclic), then this latter condition is equivalent, by Hilbert's Theorem 90, to the condition that θ not have the form $\chi \circ N_{E/F}$ for a character χ of F^\times .

(2.04) If σ is a *primitive* N -dimensional representation of $W(F)$ (that is, $\sigma \not\cong \text{Ind}_{E/F} \mu$ for $[E:F] > 1$), then there is an extension L/F which is galois and tamely ramified such that $\sigma|_{W(L)}$ is monomial [R].

Looking at the above list, it is clear that if one wishes to use the Langlands classification to describe the set of irreducible supercuspidal representations of $GL_N(F)$, one should start by attempting to find those representations which correspond to representations of the form $\text{Ind}_{E/F} \theta$, where E/F is cyclic. Historically, in the case of GL_2 , the “correct” supercuspidal representation, let us denote it by $\pi_{E/F}(\theta)$, was obtained by decomposing the “Weil representation” of $GL_2(F)$ associated to the extension E/F (see [J-L]). More recently, $\pi_{E/F}(\theta)$ has been shown to exist for general N by Kazhdan [Ka], who showed the existence of a representation whose character satisfied a certain condition, and Henniart [He1], who proved that this representation was the “right” one from the point of view of the Langlands program. An important observation here is that the existence of $\pi_{E/F}(\theta)$ followed from global arguments; in particular, even where a complete description of the set of supercuspidal representations of $GL_N(F)$ is given (notably, when N is prime [Ca] or when $p \nmid N$ [Ho1, Mo]) it is often difficult to identify $\pi_{E/F}(\theta)$ explicitly (see below).

Starting from a knowledge of the existence of $\pi_{E/F}(\theta)$ one may proceed, in certain cases, to extend this correspondence to give a Langlands parametrization for the complete set of supercuspidal representations of GL_N (see [Ku1] in the case $N = 2$, [He2] in the case $N = 3$, [Mo] in the case $p \nmid N$, [Ku-Mo] in the case that N is prime). Here, however, it is necessary to know explicit information about the set of supercuspidal representations of $GL_N(F)$. A knowledge of such properties will follow from the considerations in the next section (see (3.10) below).

3. Construction of supercuspidal representations by induction. It is reasonable on the basis of known cases to make the following

(3.01) CONJECTURE. Let G be a reductive group over a p -adic field F and let π be an irreducible supercuspidal representation of G . Then there exists a maximal open compact-modulo-center subgroup, K , of G which is unique up to conjugacy and an irreducible (smooth) representation, σ , of K which is unique up to equivalence such that $\pi \cong \text{Ind}_K^G \sigma$.

(3.02) REMARK. In this context the correct notion of induction is that of *smooth induction*: If (μ, W) is a smooth representation of an open, compact-mod-center subgroup H of G , then the representation $\text{Ind}_H^G \mu$ is defined via the action of right translation of G on the space of functions f from G to W which are compactly supported modulo center and which satisfy $f(hx) = \mu(h)(f(x))$ for h in H , x in G . It is not hard to see that $\text{Ind}_H^G \mu$ is smooth. If, in addition, $\text{Ind}_H^G \mu$ is irreducible, then $\text{Ind}_H^G \mu$ is *a fortiori* supercuspidal. This follows, first, from the fact that a coefficient of $\text{Ind}_H^G \mu$ may be obtained by extending a coefficient of μ to be 0 away from H and, second, from the fact that if a smooth representation

is irreducible, then all of its coefficients are compactly supported modulo center as soon as one is.

In order to consider whether Conjecture (3.01) holds, it is necessary to describe the structure of the set of maximal, open, compact-mod-center subgroups of G . In general, this description relies heavily on Bruhat-Tits theory [Br]. In the case of $G = \mathrm{GL}_N(F)$, which I will now consider, one is helped by the fact that $\mathrm{GL}_N(F)$ is the group of units in the algebra $A = A_N = M_N(F)$ of F -endomorphisms of N -space over F . In particular, there exist certain subrings of A which play a role in A very much analogous to that played by the ring of integers \mathcal{O}_F in F . Let us be precise. (In what follows I will rely heavily on [B-F2].)

(3.03) DEFINITION. A subring \mathcal{A} of A is called a *principal order* in A if

1. \mathcal{A} is a free \mathcal{O}_F -submodule of A of maximal rank;
2. the radical, $P_{\mathcal{A}}$, of \mathcal{A} is principal as a left or, equivalently, right ideal of \mathcal{A} .

For such a principal order \mathcal{A} , we denote the subgroup \mathcal{A}^\times of G by $U_{\mathcal{A}}$ and we set $U_{\mathcal{A}}^n = 1 + P_{\mathcal{A}}^n$ for $n \geq 1$.

(3.04) [B-F2] Let \mathcal{A} be a principal order in A and set $K_{\mathcal{A}} = N_G(U_{\mathcal{A}})$. Then $K_{\mathcal{A}}$ is a maximal, open, compact-modulo-center subgroup of G and all subgroups arise in this way.

Having described the set of maximal, open, compact-modulo-center subgroups of G , one then inquires as to what sort of property an irreducible representation σ of such a subgroup K should satisfy in order that $\mathrm{Ind}_K^G \sigma$ be irreducible. As an example, consider the case that $\mathcal{A} = M_N(\mathcal{O}_F)$ so that $U_{\mathcal{A}} = \mathrm{GL}_N(\mathcal{O}_F)$ and $K_{\mathcal{A}} = F^\times \cdot \mathrm{GL}_N(\mathcal{O}_F)$. Then the following goes back to Mautner [Ma] in the case $N = 2$.

(3.05) Let σ be an irreducible representation of $K_{\mathcal{A}} = F^\times \cdot \mathrm{GL}_N(\mathcal{O}_F)$ which is trivial on $U_{\mathcal{A}}^1$. Let $\bar{\sigma}$ be the restriction of σ to $U_{\mathcal{A}}$ viewed as a representation of $U_{\mathcal{A}}/U_{\mathcal{A}}^1 \cong \mathrm{GL}_N(\mathcal{O}_F/P_F)$. Then $\mathrm{Ind}_{K_{\mathcal{A}}}^G \sigma$ is irreducible if and only if $\bar{\sigma}$ is cuspidal in the sense of finite linear groups.

More generally, suppose that σ is an irreducible representation of some subgroup $K_{\mathcal{A}}$ of G , \mathcal{A} arbitrary, and suppose that σ has *minimal level* n , i.e., that $U_{\mathcal{A}}^n$ is contained in $\ker \sigma$ and that $U_{\mathcal{A}}^{n-1}$ is not contained in $\ker \sigma \otimes (\chi \circ \det)$ for any character χ of U_F . Suppose also that $n \geq 2$. One may then ask whether there is a property of σ , *depending only on* $\sigma|_{U_{\mathcal{A}}^{n-1}}$, which ensures that $\mathrm{Ind}_{K_{\mathcal{A}}}^G \sigma$ be irreducible.

The answer is that there is such a property, due originally to Carayol [Ca] (see also, Howe [Ho2]). To describe this property, we must first describe the topological dual, $(U_{\mathcal{A}}^{n-1}/U_{\mathcal{A}}^n)^\wedge$, of $U_{\mathcal{A}}^{n-1}/U_{\mathcal{A}}^n$.

(3.06) 1. For $r < n$, the pairing of $P_{\mathcal{A}}^r/P_{\mathcal{A}}^n \times P_{\mathcal{A}}^{1-n}/P_{\mathcal{A}}^{1-r}$ into F/P_F induced by the pairing $(x, y) \rightarrow \mathrm{tr} \, xy$ is nondegenerate.

2. Fix a character ψ of F of conductor P_F and for b in $P_{\mathcal{A}}^{1-n}/P_{\mathcal{A}}^{1-r}$ define the function $\psi_b = \psi_{(b, r, n)}$ on $P_{\mathcal{A}}^r/P_{\mathcal{A}}^n$ by $\psi_b(x) = \psi(\mathrm{tr} \, bx)$. Then the map $b \mapsto \psi_b$ induces an isomorphism of $P_{\mathcal{A}}^{1-n}/P_{\mathcal{A}}^{1-r}$ with $(P_{\mathcal{A}}^r/P_{\mathcal{A}}^n)^\wedge$.

3. For $r \geq n/2$, and b in $P_{\mathcal{A}}^{1-r}/P_{\mathcal{A}}^{1-n}$, define the function ψ_b^\times on $U_{\mathcal{A}}^r/U_{\mathcal{A}}^n$ by $\psi_b^\times(y) = \psi_b(y-1)$. Then the map $b \mapsto \psi_b^\times$ induces an isomorphism of $P_{\mathcal{A}}^{1-r}/P_{\mathcal{A}}^{1-n}$ with $(U_{\mathcal{A}}^r/U_{\mathcal{A}}^n)^\wedge$.

We now wish to pick out a set of elements b in $P_{\mathcal{A}}^{1-n}$ having the property that if a representation σ of $K_{\mathcal{A}}$ contains ψ_b^\times on restriction to $U_{\mathcal{A}}^{n-1}$, then $\text{Ind}_{K_{\mathcal{A}}}^G \sigma$ is irreducible. Call a subset S of $P_{\mathcal{A}}^{1-n}$ a *generic set of level $1-n$* if

(3.07) 1. S is invariant under conjugation by elements of $K_{\mathcal{A}}$.

2. $S + P_{\mathcal{A}}^{2-n} \subset S$.

3. If x is not in $K_{\mathcal{A}}$, then $xSx^{-1} \cap S$ is empty.

Then we have (see [Ho2])

(3.08) If S is a generic set of level $1-n$ and if σ is an irreducible representation of $K_{\mathcal{A}}$ for which $\sigma|_{U_{\mathcal{A}}^{n-1}}$ contains some ψ_b , b in S , then σ has minimal level n and $\text{Ind}_{K_{\mathcal{A}}}^G \sigma$ is irreducible.

What is left is to construct such a set.

(3.09) Let S be the set of elements b in $P_{\mathcal{A}}^{1-n}$ satisfying the following properties:

1. $E = F[b]$ is a subfield of A of degree N over F .

2. E^\times lies in $K_{\mathcal{A}}$.

3. b is *minimal* in E , that is, $(\nu_E(b), e(E/F)) = 1$ and

$$\mathcal{O}_{E_{nr}} = \mathcal{O}_F[(N_{E/E_{nr}} b)/\bar{\omega}_F^{1-n}],$$

where E_{nr}/F is the unramified part of E/F and where $\bar{\omega}_F$ is any prime element of \mathcal{O}_F .

Then S is the unique maximal generic set of level $1-n$.

(3.10) REMARKS. 1. For the construction of generic sets for other reductive groups see [Hi] and [Mr].

2. Call an irreducible representation σ of $K_{\mathcal{A}}$ *very cuspidal of level n* if either $n = 1$ and σ is one of the representations considered in (3.05) or if $n \geq 2$ and the restriction of σ to $U_{\mathcal{A}}^{n-1}$ contains a character ψ_b with b generic of level $1-n$. Then in case N is prime, every irreducible supercuspidal representation of G has the form $\text{Ind}(\sigma \otimes (\chi \circ \det))$ for some very cuspidal representation σ [Ca]. This is *not* the case if N is composite, a situation which will be discussed in §4.

3. Suppose, to avoid technical difficulties, that $n = 2m$ is even. Then very cuspidal representations are quite easy to construct. In fact, the restriction of such a representation σ to the subgroup $U_{\mathcal{A}}^m$ will contain a representation ψ_b , where b satisfies the conditions of (3.09). Furthermore, the stability subgroup of ψ_b in $K_{\mathcal{A}}$ will be the group $E^\times U_{\mathcal{A}}^m$. Thus, σ will have the form $\text{Ind } \theta \psi_b$, where θ is a character of E^\times whose restriction to $U_E^m = U_{\mathcal{A}}^m \cap E^\times$ coincides with the restriction of ψ_b . Since it is easy to see that if E/F is cyclic, then $\theta^\tau \neq \theta$ for $\tau \neq 1$ in $\text{Gal}(E/F)$, one is led to ask whether $\text{Ind } \sigma$ is in fact Kazhdan's representation $\pi_{E/F}(\theta)$. The answer to this is of deep significance to the arithmetic of local fields. It is that, *if E/F is tamely ramified*, then there is a character ω of E^\times , independent of θ , so that $\text{Ind } \sigma = \pi_{E/F}(\theta\omega)$ ([Mo]; see also, [B-F1] for the case

of a division algebra), while if E/F is wildly ramified (so that, in particular, $p \mid N$), then, in general, $\text{Ind } \sigma$ need not have the form $\pi_{E_1/F}(\theta_1)$ for any pair (E_1, θ_1) and, even if $\text{Ind } \sigma = \pi_{E_1/F}(\theta_1)$, then it may be the case that $E_1 \neq E$ (see [Ku2] for a complete description of this phenomenon in case $N = 2$).

4. If L/F is a tamely ramified extension, then the construction of supercuspidal representations by induction from very cuspidal representations can be used to define a “lift” of representations from $\text{GL}_N(F)$ to $\text{GL}_N(L)$. It is this lift which enables one to extend the map $\theta \mapsto \pi_{E/F}(\theta)$ to a Langlands parametrization in the case that N is prime [Ku-Mo].

4. The case that N is composite. Since the set of representations constructed by Carayol does not account for all irreducible supercuspidal representations in case N is composite, it may be helpful to describe the kind of representation of $W(F)$ which ought to parametrize Carayol’s representations.

(4.01) DEFINITION. Call an irreducible representation σ of $W(F)$ *degenerate* if there is an extension L/F , a representation τ of $W(L)$ of degree greater than one, and a character χ of L^\times such that

1. $\sigma = \text{Ind}_{L/F} \tau$ and
2. $f(\tau \otimes \chi) < f(\sigma)$,

where $f(\sigma)$ is the Artin conductor of σ .

In particular, if $\sigma = \text{Ind}_{E/F} \theta$, where E/F is cyclic, then σ is nondegenerate if and only if for all extensions L/F intermediate to E/F and with $L \neq E$ and for all characters χ of L^\times , we have that $f(\theta \otimes \chi \circ N_{E/L}) \geq f(\theta)$.

Let us look at this in a slightly different light. Suppose that $f(\theta) \geq 2$ and fix a character ψ of F ; set $\psi_{E/F} = \psi \circ \text{Tr}_{E/F}$. Then (see (3.06)) there is an element α in E such that, for all x in $U_E^{f(\theta)-1}$, we have $\theta(x) = \psi_{E/F}(\alpha(x-1))$. Then $\sigma = \text{Ind}_{E/F} \theta$ is nondegenerate if and only if, given any element β in E such that $F[\beta] \subsetneq E$, we have that $\nu_E(\alpha - \beta) \leq \nu_E(\alpha)$. It is important to note that this condition on α is in fact equivalent to the minimality of α in E (see (3.09)) if E/F is tame and that the condition given in (3.09) is only slightly stronger (for technical reasons) in case E/F has wild ramification. So in sum,

(4.02) We expect nondegenerate representations of $W(F)$ to parametrize the set of supercuspidals constructed by Carayol.

Indeed, in case $p \nmid N$, where we have complete information [Mo], this expectation is fulfilled.

The above considerations should (I hope!) motivate what follows (see [Ku-Md]). Write $N = RS$, $R \neq 1$, $S \neq 1$, and fix an extension E/F of degree R . Since it is easier to work geometrically, I will fix a vector space V of dimension S over E and will write $V_{E/F}$ for the F -vector space obtained from V by restriction of scalars. Then $A_E = \text{End}_E(V)$ is naturally a subring of $A_F = \text{End}_F(V_{E/F})$ and if I fix a principal order \mathcal{A}_E in A_E there is a unique principal order \mathcal{A}_F of A_F with the property that $P_{\mathcal{A}_F}^n \cap A_E = P_{\mathcal{A}_E}^n$ for all n . Fix, as always, a character ψ of F having conductor P_F and fix also a minimal element α of E/F . Then α will be, of course, a scalar in A_E and its \mathcal{A}_E -valuation (in the obvious sense),

$\nu_{\mathcal{A}_E}(\alpha)$, will be equal to $e(\mathcal{A}_E) \cdot \nu_E(\alpha)$, where $e(\mathcal{A}_E)$ is the *ramification degree* of \mathcal{A}_E (see [B-F2]). If we then set $n = 1 - \nu_{\mathcal{A}_E}(\alpha)$ and pick α so that $n \geq 2$, we obtain in this way a character ψ_α on $U_{\mathcal{A}_F}^{n-1}/U_{\mathcal{A}_F}^n$ which is definitely not minimal.

Now, the idea is to fix α and to consider characters of the form $\psi_{\alpha+\beta}$ on, say, $U_{\mathcal{A}_F}^{r-1}/U_{\mathcal{A}_F}^r$, $n-1 < r \leq n/2+1$, where β lies on $P_{\mathcal{A}_F}^{1-r}/P_{\mathcal{A}_F}^{2-r}$. (Note: Strictly speaking, one may have to consider subgroups of the form $U_{\mathcal{A}_E}^{r-1}U_{\mathcal{A}_F}^m$ with $n-1 \leq m \leq r$. This causes additional difficulties since such subgroups are not normal in $K_{\mathcal{A}_F}$. I will not discuss this further here.) That is, one should consider characters of $U_{\mathcal{A}_F}^r$ which *lie over* ψ_α . In case E/F is tamely ramified, it is a fundamental fact that β may, after possibly conjugating by an element of $K_{\mathcal{A}_F}$, be taken to lie in \mathcal{A}_E [Ho1]. This is the starting point of a reduction of rank argument that produces the other supercuspidals (see also [H-M]). In fact, upon a closer examination of this situation, one sees that it is only the restriction of ψ_b to $U_{\mathcal{A}_E}^r$ which really matters. Let me be more precise.

We have the following diagram

$$\begin{array}{ccc} (U_{\mathcal{A}_F}^{r-1}/U_{\mathcal{A}_F}^r)^\wedge & \cong & P_{\mathcal{A}_F}^{1-r}/P_{\mathcal{A}_F}^{2-r} \\ \text{restriction } \downarrow & & \downarrow \\ (U_{\mathcal{A}_E}^{r-1}/U_{\mathcal{A}_E}^r)^\wedge & \cong & P_{\mathcal{A}_E}^{1-r}/(P_{\mathcal{A}_E}^{r-1} + P_{\mathcal{A}_F}^r)^* \end{array}$$

where, in general, Q^* is the set of elements x in A_F for which $\text{tr}(xQ) \subset P_F$. Now it is easy to see that $(P_{\mathcal{A}_E}^{r-1} + P_{\mathcal{A}_F}^r)^* = (\text{Im } A_\alpha \cap P_{\mathcal{A}_F}^{1-r}) + P_{\mathcal{A}_F}^{2-r}$ where A_α is the map from A_F to A_F defined by $A_\alpha(x) = \alpha x \alpha^{-1} - x$. Thus restricting ψ_β to $U_{\mathcal{A}_E}^{r-1}$ amounts to considering the image of β in $P_{\mathcal{A}_F}^{1-r}/(\text{Im } A_\alpha \cap P_{\mathcal{A}_F}^{1-r}) + P_{\mathcal{A}_F}^{2-r}$. Clearly, if we are to work meaningfully with the character $\psi_{\alpha+\beta}$ it is necessary to identify the space $P_{\mathcal{A}_F}^{1-r}/(\text{Im } A_\alpha \cap P_{\mathcal{A}_F}^{1-r}) + P_{\mathcal{A}_F}^{2-r}$ with some less abstract space. In case E/F is tame, this is easy: in fact, $P_{\mathcal{A}_F}^j = (\text{Im } A_\alpha \cap P_{\mathcal{A}_F}^j) \oplus P_{\mathcal{A}_E}^j$ (as \mathcal{A}_E -modules) for all j and so we get an \mathcal{A}_E -module isomorphism of $P_{\mathcal{A}_F}^{1-r}/(\text{Im } A_\alpha \cap P_{\mathcal{A}_F}^{1-r}) + P_{\mathcal{A}_F}^{2-r}$ with $P_{\mathcal{A}_E}^{1-r}/P_{\mathcal{A}_E}^{2-r}$. It is this isomorphism that enables one to use reduction on rank.

In case E/F is wild we have no splitting of $P_{\mathcal{A}_F}^j$ of the sort described above. Nonetheless, we have the following:

(4.03) Let f be the irreducible polynomial of α over F and set $g(x) = f(\alpha+x)$. Then $g(x)$ may be viewed as a function on A_F and we define the function S_α from A_F to A_F to be the “linear term” (in the obvious sense) of $\alpha^{1-R}g(x)$.

(4.04) The sequence

$$P_{\mathcal{A}_F}^j \xrightarrow{A_\alpha} P_{\mathcal{A}_F}^j \xrightarrow{S_\alpha} P_{\mathcal{A}_F}^j \xrightarrow{A_\alpha} P_{\mathcal{A}_F}^j$$

is an exact sequence of \mathcal{A}_E -modules for all j .

As a corollary we get

(4.05) S_α induces an \mathcal{A}_E -isomorphism, $\overline{S}_{\alpha,r}$, of

$$P_{\mathcal{A}_F}^{1-r}/(\text{Im } A_\alpha \cap P_{\mathcal{A}_F}^{1-r}) + P_{\mathcal{A}_F}^{2-r} \quad \text{with} \quad P_{\mathcal{A}_E}^{1-r}/P_{\mathcal{A}_E}^{2-r} \quad \text{for all } r.$$

Thus one may identify the dual of $U_{A_E}^{r-1}/U_{A_E}^r$ with $P_{A_E}^{1-r}/P_{A_E}^{2-r}$, via the map $\overline{S}_{\alpha,r}$, in such a way that one may begin a reduction of rank argument. In particular one may show in this way that

(4.06) [Ku-Md] Conjecture (3.01) holds for $GL_4(F)$.

REFERENCES

- [Bo] A. Borel, *Automorphic L-functions*, Automorphic Forms, Representations, and L-Functions, Proc. Sympos. Pure Math., vol. 33, part 2, Amer. Math. Soc., Providence, R.I., 1979, pp. 27–63.
- [Br] F. Bruhat, *Sur une classe de sous-groupes compacts maximaux des groupes de Chevalley sur un corps p-adique*, Inst. Hautes Études Sci. Publ. Math. **23** (1964), 46–74.
- [B-F1] C. J. Bushnell and A. Frohlich, *Gauss sums and p-adic division algebras*, Lecture Notes in Math., vol. 987, Springer-Verlag, 1983.
- [B-F2] —, *Non-abelian congruence Gauss sums and p-adic simple algebras*, Proc. London Math. Soc. (3) **50** (1985), 207–264.
- [Ca] H. Carayol, *Représentations cuspidales du groupe linéaire*, Ann. Sci. École Norm. Sup.
- [Cs] W. Casselman, *Introduction to the theory of admissible representations of p-adic reductive groups*, Preprint.
- [Ct] P. Cartier, *Sur les représentations des groupes réductifs p-adiques et leurs caractères*, Sem. Bourbaki 1975/76, exp. 471, Lecture Notes in Math., vol. 567, Springer-Verlag, 1977.
- [D] P. Deligne, *Formes modulaires et représentations de $GL(2)$* , Modular Functions of One Variable. II, Lecture Notes in Math., vol. 349, Springer-Verlag, 1973, pp. 501–597.
- [HC] Harish-Chandra, *Eisenstein series over finite fields*, Functional Analysis and Related Fields, Springer-Verlag, 1970, pp. 76–88.
- [He1] G. Henniart, *On the local Langlands conjecture for $GL(n)$: The cyclic case*, Ann. of Math. (2) **123** (1986), 145–203.
- [He2] —, *La conjecture de Langlands locale pour $GL(3)$* , Mém. Soc. Math. France (N.S.) **11/12** (1984).
- [Hi] H. Hijikata, *Some supercuspidal representations induced from parahoric subgroups* (Taniguchi Sympos., Katata 1983), Birkhauser, 1984.
- [Ho1] R. Howe, *Tamely ramified supercuspidal representations of GL_n* , Pacific J. Math. **73** (1977), 437–460.
- [Ho2] —, *Some qualitative results on the representation theory of GL_n over a p-adic field*, Pacific J. Math. **73** (1977), 479–538.
- [H-M] R. Howe and A. Moy, *Harish-Chandra homomorphisms for p-adic groups*, CBMS Regional Conf. Ser. in Math., no. 59, Amer. Math. Soc., Providence, R.I., 1985.
- [J] H. Jacquet, *Représentations des groupes linéaires p-adiques*, Theory of Group Representations and Harmonic Analysis (C.I.M.E., II Ciclo, Montecatini, 1970), Edizioni Cremonese, Roma, 1971, pp. 119–220.
- [J-L] H. Jacquet and R. P. Langlands, *Automorphic forms on $GL(2)$* , Lecture Notes in Math., vol. 114, Springer-Verlag, 1970.
- [Ka] D. Kazhdan, *On lifting*, Lie Groups Representations. II, Lecture Notes in Math., vol. 1024, Springer-Verlag, 1983.
- [Ku1] P. C. Kutzko, *The Langlands conjecture for GL_2 of a local field*, Ann. of Math. (2) **112** (1980), 381–412.
- [Ku2] —, *The exceptional representations of GL_2* , Compositio Math. **51** (1984), 3–14.
- [Ku-Md] P. C. Kutzko and D. Manderscheid, *On the supercuspidal representations of GL_4* , Duke Math. J. **52** (1985), 841–867.
- [Ku-Mo] P. C. Kutzko and A. Moy, *On the local Langlands conjecture in prime dimension*, Ann. of Math. (2) **121** (1985), 495–516.
- [Ma] F. Mautner, *Spherical functions over p-adic fields*. II, Amer. J. Math. **86** (1964), 171–200.

[Mo] A. Moy, *Local constants and the tame Langlands correspondence*, Amer. J. Math. (to appear).

[MR] L. Morris, *Filtrations and supercuspidal representations*, Preprint.

[R] J. F. Rigby, *Primitive linear groups containing a normal nilpotent subgroup larger than the centre of the group*, J. London Math. Soc. **35** (1960), 389–400.

[S] A. Silberger, *Introduction to harmonic analysis on reductive p -adic groups*, Math. Notes, no. 23, Princeton Univ. Press, Princeton, N.J., and Univ. of Tokyo Press, Tokyo, 1979.

UNIVERSITY OF IOWA, IOWA CITY, IOWA 52240, USA

Integrable Lattice Models and Branching Coefficients

TETSUJI MIWA

This is a brief survey of the recent work [1–4] in Kyoto on integrable lattice models and their connection to the representation theory of the affine Lie algebras. Since the topic is related to two different areas, I have not hesitated to explain some well-known facts as far as they are necessary for the description of our results.

In §1 we state Onsager-Yang's result on the magnetization of the two-dimensional Ising model. The point is that it is related to modular forms. In §2 we present a hierarchy of integrable lattice models. They are N -state models with the nearest neighbor interactions satisfying the star-triangle relation. The Ising model is just the case $N = 2$. In §3 the local height probabilities (LHP) for these models are shown to be related to the branching coefficients appearing in the representation theory of the affine Lie algebra $A_1^{(1)}$. In §4 a parallel result is given for another hierarchy of integrable models. It is related to the decomposition of the tensor product of the $A_1^{(1)}$ modules. The last section is devoted to discussion.

1. The magnetization of the two-dimensional Ising model. The Ising model is a lattice model of a magnet. It is exactly solvable if the dimension of the lattice is less than or equal to two. Since the two-dimensional case exhibits a critical singularity, it has been affording the simplest test run for several theories in statistical mechanics and quantum field theory.

Consider a two-dimensional square lattice \mathcal{L} of size $m \times m$. To each site i of \mathcal{L} we associate a variable $\sigma(i)$ which takes the value $+1$ or -1 . We determine the interaction of the spins in the microscopic level as follows. Consider the following sum $Z(m)$ called the partition function

$$Z(m) = \sum_{\sigma} \prod^{(1)} W(1, \sigma(i), \sigma(j)) \prod^{(2)} W(2, \sigma(i), \sigma(j)).$$

The quantity $W(k, a, b)$ ($k = 1, 2$; $a, b = +1, -1$) is called the Boltzmann weight. The product $\prod^{(k)}$ is over all the vertical pairs i, j if $k = 1$ and over all the horizontal pairs i, j if $k = 2$. The sum \sum_{σ} is over all possible states of the spins. We suppose that the probability of the occurrence of a particular state σ is given

by

$$P(m, \sigma) = \prod^{(1)} W(1, \sigma(i), \sigma(j)) \prod^{(2)} W(2, \sigma(i), \sigma(j)) / Z(m).$$

The particular case where the Boltzmann weight is of the form

$$W(k, a, b) = \exp(ab \log(z(k))/2)$$

is called the Ising model. Here $z(1)$ and $z(2)$ are the parameters related to the temperature and the anisotropy of the lattice. We note the symmetries

$$W(k, a, b) = W(k, b, a), \quad W(k, a, b) = W(k, -a, -b). \quad (1)$$

Let us denote by $\sigma(0)$ the spin variable located at the center of \mathcal{L} . The macroscopic quantity we are interested in here is the following expectation value in the infinite limit of m :

$$\sum_{\sigma} \sigma(0) P(m, \sigma). \quad (2)$$

We need a more precise definition of this quantity.

Because of the symmetry (1) the sum (2) is zero. We must consider the partial sum with the boundary spins being fixed to $+1$. The precise meaning of this is as follows. If we restrict the parameters to be in the low temperature region $z(k) \gg 1$, the largest contribution to (2) comes from two "ground" states where the spins take the same values $+1$ or -1 . Throw away the latter state and consider the partial sum in (2) only over those states for which the number of -1 spins is small, say less than m' . The fact is that if $m \gg m' \gg 1$ this partial sum is independent of m and m' up to higher-order terms in $1/z(1)$ and $1/z(2)$. Thus we obtain a well-defined series in $1/z(1)$ and $1/z(2)$ by letting first m go to infinity and second m' go to infinity. This is called the magnetization. We denote it by $M(0)$. The nonvanishing of $M(0)$ in the low temperature region is called the spontaneous break-down of the symmetry (1).

Consider the following parametrization of $z(1)$ and $z(2)$ in terms of the Jacobian theta functions (quoted from [5])

$$\begin{aligned} z(1) &= (\sqrt{k'} H_1(iu) - iH(iu)) / \sqrt{k} \Theta(iu), \\ z(2) &= i(\sqrt{k'} \Theta_1(iu) + \Theta(iu)) / \sqrt{k} H(iu). \end{aligned} \quad (3)$$

Then $M(0)$ is independent of u and is given by

$$(1 - k^2)^{1/8} = \prod_{n=1}^{\infty} (1 - q^{2n-1}) / (1 + q^{2n-1}),$$

where $q = \exp(-\pi K'/K)$ and K and K' are the half-period magnitudes with the elliptic modulus k . This is the result of Onsager and Yang. It exhibits a singularity at $k = 1$. This is called the critical singularity. We note that the parametrization (3) becomes trigonometric at criticality.

Before proceeding further let us give one word for $M(0)$ in the high temperature region. It is just zero whatever is the choice of the boundary spins. In other words the symmetry (2) is not broken in this region.

2. The N -state generalization of the Ising model. One can split Onsager-Yang's result on $M(0)$ into two pieces:

- (i) $M(0)$ depends only on k ,
- (ii) $M(0)$ is given by a simple infinite product in q .

First we try to explain the situation corresponding to (i) in general N -state models. The part (ii) will be discussed later.

The mechanism for the validity of (i) was unveiled by Baxter. Moreover, he invented the method of corner transfer matrix (CTM) for the computation of such quantities similar to the magnetization. His method is applicable to general integrable models.

The model is integrable if the Boltzmann weight satisfies the so-called star-triangle relation (STR). Here we avoid going into details on STR, since it is not necessary for the main line of this review. We concentrate on explaining the characteristic feature of the solutions to STR obtained in [1].

The idea is to generalize the model in such a way that the number of states of each spin is $N > 2$. In [12] Fateev and Zamolodchikov found such solutions for arbitrary N . The Boltzmann weights of their solutions are parametrized by trigonometric functions. Thus the case $N = 2$ merely corresponds to the critical Ising model.

Some perturbative experiments from their solutions by using FORTRAN and MACSYMA suggest the existence of N irreducible extensions. After a detailed computation for $N = 3$ we reached the parametrization of the Boltzmann weight for general N . It is a certain product of $2N$ Jacobian theta functions with the arguments u and k [1].

The first obvious remark is that the number of parameters does not increase for higher N . Thus, contrary to the case $N = 2$, the integrable cases for higher N constitute rather complicated submanifolds in the higher-dimensional space of the Boltzmann weights. This is a defect if we want to analyze the general nature of the N -state Ising model. But the integrable cases with the elliptic parametrization are certainly not critical; thus the local height probabilities (see below) are well worth the investigation.

Let us describe our solutions in a little bit more detail. We call the variable at each site the height variable and denote it by $l(i)$. We assume that $l(i)$ takes its value in Z_n . There are three different cases classified by s and n :

$$s = 0, \quad N = 2n, \quad s = 0, \quad N = 2n + 1, \quad s = 1, \quad N = 2n.$$

The symmetry (1) is modified to

$$W(k, a, b) = W(k, -a - s, -b - s). \quad (4)$$

Let us define the local height probabilities (LHP). Instead of (2) we consider the partial sum

$$\sum_l P(m, l)$$

with the center spin being fixed to a and the boundary spins being fixed to b (the precise meaning of this is just as in the case of the magnetization). We obtain a

series in $q = \exp(-4\pi K/NK')$. The low temperature limit corresponds to $q = 0$ and the critical limit corresponds to $q = 1$. We denote the LHP by $P(a, b)$. The point is that the symmetry (4) is not broken. Thus without loss of generality we can restrict to $a, b = 0, 1, \dots, n - s$.

By applying the corner transfer matrix (CTM) we obtain the following expression for LHP:

$$P(a, b) = \lim E(-q^{(2a+s)/2}, q^{N/2})F(m, a, b)/M(m, b). \quad (5)$$

Here $M(m, b)$ is the normalizing factor:

$$M(m, b) = \sum_{a=0}^{n-s} \#(2a + s, N)E(-q^{(2a+s)/2}, q^{N/2})F(m, a, b),$$

where

$$\#(j, N) = \begin{cases} 1 & \text{if } j \equiv 0 \pmod{N}, \\ 2 & \text{if } j \not\equiv 0 \pmod{N}, \end{cases}$$

and

$$\begin{aligned} E(z, q) &= \prod_{j=1}^{\infty} (1 - zq^{j-1})(1 - z^{-1}q^j)(1 - q^j), \\ F(m, a, b) &= \sum_{x, y} q^{H(m, x, y)}, \\ H(m, x, y) &= \sum_{i=1}^m (i - 1/2)|x(i) - x(i + 1)| + \sum_{i=2}^m (i - 1)y(i). \end{aligned}$$

The sum in $H(m, x, y)$ is over $x(2), \dots, x(m) = 0, 1, \dots, n - s$ (while we keep $x(1) = a$ and $x(m + 1) = b$) and $y(2), \dots, y(m) = 0, 1$ with the exception that if $2x(i) + s \equiv 0 \pmod{N}$ then $y(i) = 0$ only.

Now it is a good chance to discuss (ii), the automorphic property. In general it is not true that LHP is expressed as a simple infinite product. Although it is not clear from the above expression (5), the fact is that $P(a, b)$ is related to modular forms. In fact, in §3 we relate it to the character formulas for the $A_1^{(1)}$ modules. Then the automorphic property of $P(a, b)$ follows immediately. Because of this property we can compute the singular behavior of $P(a, b)$ at the criticality.

3. The affine Lie algebra $A_1^{(1)}$ and the branching coefficients. Now we change our topic to representation theory, the character formula, and the related theta function identities. These are all that we need. So we take the shortest cut with no sophistication.

The affine Lie algebra $A_1^{(1)}$ is the central extension of $\mathfrak{sl}(2, C) \otimes C[t, 1/t]$ with the Lie bracket

$$[X \otimes t^i, Y \otimes t^j] = [X, Y] \otimes t^{i+j} + i\delta_{i, -j} \text{trace } XYc.$$

It is decomposed into three pieces:

$$A_1^{(1)} = n_+ + h + n_-,$$

where n_+ is spanned by $X \otimes t^m$ ($m > 0$) and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, n is spanned by $X \otimes t^m$ ($m < 0$) and $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, and h is spanned by $h_1 = \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}$ and c .

We denote by $V(l, k)$ (l, k : integer, $0 \leq k \leq l$) the highest weight module with the highest weight $(l-k)\Lambda_0 + k\Lambda_1$. This means that $V(l, k)$ is an irreducible $A_1^{(1)}$ module with the highest weight vector $|l, k\rangle$ satisfying

$$n_+|l, k\rangle = 0, \quad c|l, k\rangle = l|l, k\rangle, \quad h_1|l, k\rangle = k|l, k\rangle.$$

The integer l is called the level.

The vectors in $V(l, k)$ are produced from $|l, k\rangle$ by applying $f_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes (1/t)$ and $f_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. More precisely, we have the decomposition of $V(l, k)$:

$$V(l, k) = \sum_{i,j} V(i, j; l, k),$$

$$V(i, j, l, k) = \sum C f_{m_1} \cdots f_{m_{i+2j}} |l, k\rangle,$$

where the sum is over such m_1, \dots, m_{i+2j} that the number of f_0 is equal to j and the number of f_1 is equal to $i+j$.

We define the character $\text{ch}(l, k; z, q)$ of $V(l, k)$ by the following series in z and q :

$$\text{ch}(l, k; z, q) = z^{-1/2} \sum_{i,j} \dim V(i, j; l, k) z^i q^j.$$

We now follow Kac-Peterson's work [6] to obtain a certain theta function identity. Define elliptic theta functions $F(l, k; z, q)$ and $G(l, k; z, q)$ by

$$F(l, k; z, q) = \sum_{j \in \mathbb{Z}} q^{lj^2 + kj} (z^{-lj - k/2} - z^{lj + k/2}),$$

$$G(l, k; z, q) = \sum_{j \in \mathbb{Z}} q^{lj^2 + kj} (z^{-lj - k/2} + z^{lj + k/2}).$$

The character formula tells us that

$$\text{ch}(l, k; z, q) = z^{1/2} F(l, k; z, q) / E(z, q).$$

The affine Weyl group W acts on $V(l, k)$. Fix a W -orbit. Then the partial sum in $\text{ch}(l, k; z, q)$ restricted to this orbit gives rise to

$$\sharp(j, l) q^i G(l, j; z, q) / 2$$

for certain i and j . We note that $0 \leq j \leq l$ and $j \equiv k \pmod{2}$. Therefore the character $\text{ch}(l, k; z, q)$ must be decomposed into the following form:

$$\text{ch}(l, k; z, q) = \sum_j \sharp(j, l) G(l, j; z, q) b(l, j, k; q)$$

with suitable q -series $b(l, j, k; q)$. In conclusion we obtain the following theta function identity:

$$F(l+2, k+1; z, q) / E(z, q) = z^{-1/2} \sum_j \sharp(j, l) G(l, j; z, q) b(l, j, k; q).$$

This is the result of Kac and Peterson. Let

$$E(l, j, k; q) = b(l, j, k; q) \varphi(q)$$

where $\varphi(q) = \prod(1 - q^j)$. We call $E(l, j, k; q)$ the branching coefficient [7]. By using the expression for $E(l, j, k; q)$ given in [7] we obtain

PROPOSITION.

$$\lim_{m \rightarrow \infty} F(m, a, b) = q^{(b-a)/2} E(N, 2a + s, 2b + s).$$

Finally, the identities

$$\begin{aligned} F(N, k; q^{1/2}, q) &= q^{-k/4} E(q^{k/2}, q^{N/2}), \\ G(N, k; q^{1/2}, q) &= q^{-k/4} E(-q^{k/2}, q^{N/2}) \end{aligned}$$

lead us to the simple expression for LHP.

THEOREM [2].

$$P(a, b) = \frac{q^{(b-a)/2} E(-q^{(2a+s)/2}, q^{N/2}) E(q^{1/2}, q) E(N, 2a + s, 2b + s; q)}{E(q^{(2b+s+1)/2}, q^{(N+2)/2}) \varphi(q)}.$$

4. The RSOS models and the decomposition of the tensor product.

The models considered in §3 had the interactions associated to each bond of the lattice. Baxter's formalism applies to a little bit more general case: the interactions round a face.

In [8] Andrews, Baxter, and Forrester introduced a hierarchy of models and computed LHP. They called the models restricted solid on solid models (RSOS). Their models are classified by integers $L \geq 4$. The simplest case $L = 4$ is equivalent to the Ising model.

In [3] we introduced another integer parameter N and constructed RSOS models in such a way that $N = 1$ corresponds to [8]. Let us describe our models. We denote again by $l(i)$ the height variable associated to a site i . We impose the following two restrictions on a nearest pair $l(i)$ and $l(j)$:

- (i) $(l(i) + l(j) - N)/2 = 1, \dots, L - N - 1$,
- (ii) $(l(i) - l(j) + N)/2 = 0, \dots, N$.

They imply the following:

- (iii) $l(i) = 1, \dots, L - 1$.

Let $l(1), l(2), l(3), l(4)$ be the height variables round a face. The interaction between them is given by the Boltzmann weight $W(l(1), l(2), l(3), l(4))$. It is parametrized by u and k as a sum of the product of N Jacobian elliptic theta functions [3].

There are four different regimes in the parametrization space of u and k . Most recently we have completed the computation of LHP in the so-called regime III where

$$0 < \exp(-\pi K'/K) < 1, \quad -1 < u < 0.$$

The result is related to the decomposition of the tensor product of the $A_1^{(1)}$ modules.

The ground state in the regime III is specified by a pair of integers b and c satisfying the restrictions (i) and (ii). Let us denote by $P(a, b, c)$ the LHP

of the center height being equal to a with the boundary spins being b and c alternatively. The precise meaning of this is about the same as in §§1 and 2 except for the following side remark: We assume that $a \equiv b \pmod{2}$ and the parity of the length from the central site to the boundary site with the height b is even.

Consider the tensor product $V(l_1, k_1) \otimes V(l_2, k_2)$ of two $A_1^{(1)}$ modules $V(l_1, k_1)$ and $V(l_2, k_2)$. This is completely reducible as an $A_1^{(1)}$ module. The irreducible decomposition of $V(l_1, k_1) \otimes V(l_2, k_2)$ gives rise to the following theta function identity:

$$\begin{aligned} & F(l_1 + 2, k_1 + 1; z, q) F(l_2 + 2, k_2 + 1; z, q) / E(z, q) \\ &= z^{-1/2} \sum_k F(l_1 + l_2 + 2, k + 1; z, q) E(l_1, l_2, k, k_1, k_2; q). \end{aligned}$$

We call $E(l_1, l_2, k, k_1, k_2; q)$ the branching coefficient. We exploit this identity with

$$\begin{aligned} l_1 &= L - N - 2, & k_1 &= (b + c - N)/2 - 1, \\ l_2 &= N, & k_2 &= (b - c + N)/2, \\ l_1 + l_2 &= L - 2, & k &= a - 1. \end{aligned}$$

The meaning of the somewhat tricky restrictions (i), (ii), (iii) is now quite clear. Our result is the following.

THEOREM [4].

$$P(a, b, c) = \frac{q^{(b-a)/4} E(q^{a/2}, q^{L/2}) E(q^{1/2}, q) E(l_1, l_2, k, k_1, k_2; q)}{E(q^{(b+c-N)/4}, q^{(L-N)/2}) E(q^{(b-c+N+2)/4}, q^{(N+2)/2})}.$$

5. Discussion. We pursued the following steps:

- (a) to solve STR and to find the parametrization of the Boltzmann weight;
- (b) to exploit the method of corner transfer matrix and to express LHP as a one-dimensional partition sum;
- (c) to identify LHP with the branching coefficient for some representation of the affine Lie algebra.

As for (a) the natural question is: Can one exhaust solutions to STR? In this review we have discussed models with elliptic parametrization with interactions associated to either a bond or a face. There is another type of model: interactions associated with a vertex. The condition for integrability for it is called the Yang-Baxter equation (YBE).

An extensive study has been made in Leningrad on two-dimensional integrable systems based on YBE. It is called the quantum inverse scattering method (QISM). Their work covers many aspects of integrability: classical-quantum, lattice-continuum. One reason for their success is that YBE is rather directly connected to Lie algebras. A systematic search has been made for solutions to YBE along that line [9]. Roughly speaking the solutions are classified by two axes, rank M and level N .

The third axis L is supplied by the RSOS model. Therefore we are led to a more specific question:

PROBLEM 1. Fill up the three dimensions L, M, N with the integrable models.

Note that the N -state Ising model is still out of this picture. The common feature in the two examples given here is the remarkable connection between LHP and the branching coefficients: (b) and (c). To put it differently, they are somehow related to a pair of Lie algebras from which the branching coefficients come. So far, the link is only in a phenomenological sense. The answer is strongly requested to the following:

PROBLEM 2. Construct a theory which explains this link.

Finally, the relation to the conformal field theory should be mentioned. As Huse [10] pointed out, the RSOS model ($N = 1$) corresponds to the minimal theory of Belavin, Polyakov, and Zamolodchikov [11]. In [2] we have shown a similar correspondence between the N -state Ising model and the Fateev-Zamolodchikov model [12]. Therefore third we put

PROBLEM 3. Construct the conformal field theory corresponding to the RSOS models ($N > 1$).

ACKNOWLEDGMENT. Closing this review at this point, I thank Alberto Grunbaum for offering me his office during my preparation of this manuscript.

REFERENCES

1. M. Kashiwara and T. Miwa, *A class of elliptic solutions to the star-triangle relation*, Nuclear Physics **B275** [FS17] (1986), 121–134.
2. M. Jimbo, T. Miwa, and M. Okado, *Solvable lattice models with broken Z_n symmetry and Hecke's indefinite modular forms*, Nuclear Physics **B275** [FS17] (1986), 517–545.
3. E. Date, M. Jimbo, T. Miwa, and M. Okado, *Fusion of the eight vertex SOS model*, Lett. Math. Phys. **12** (1986), 209–215.
- 4a. —, *Automorphic properties of local height probabilities for integrable solid-on-solid models*, Phys. Rev. B, Rapid Comm. **35** (1987), 2105–2107.
- 4b. E. Date, M. Jimbo, A. Kuniba, T. Miwa, and M. Okado, *Exactly solvable SOS models: Local height probabilities and theta function identities*, Preprint, RIMS, Kyoto Univ., 1987.
- 4c. —, *Exactly solvable SOS models II: Star-triangle relation and combinatorial identities*, Preprint, RIMS, Kyoto Univ., 1987.
5. R. J. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, New York, 1982.
6. V. G. Kac and D. H. Peterson, *Infinite-dimensional Lie algebras, theta functions and modular forms*, Adv. in Math. **53** (1984), 125–264.
- 7a. M. Jimbo and T. Miwa, *Irreducible decomposition of fundamental modules for $A_l^{(1)}$ and $C_l^{(1)}$, and Hecke modular forms*, Adv. Stud. Pure Math. **4** (1984), 97–119.
- 7b. —, *On a duality of branching rules for affine Lie algebras*, Adv. Stud. Pure Math. **6** (1985), 17–65.
8. G. E. Andrews, R. J. Baxter, and P. J. Forrester, *Eight-vertex SOS model and generalized Rogers-Ramanujan-type identities*, J. Statist. Phys. **35** (1984), 193–266.
- 9a. A. A. Belavin and V. G. Drinfeld, *Solutions of the classical Yang-Baxter equation for simple Lie algebras*, Funktsional Anal. Appl. **16** (1982), 159–180.
- 9b. P. P. Kulish, N. Yu. Reshetikhin, and E. K. Sklyanin, *Yang-Baxter equation and representation theory. I*, Lett. Math. Phys. **5** (1981), 393–403.

- 9c. V. G. Drinfeld, *Hopf algebras and the quantum Yang-Baxter equation*, Dokl. Akad. Nauk SSSR **283** (1985), 254–258.
- 9d. M. Jimbo, *A q -difference analogue of $U(g)$ and the Yang-Baxter equation*, Lett. Math. Phys. **10** (1985), 63–69.
- 10. D. A. Huse, *Exact exponents for infinitely many new multicritical points*, Phys. Rev. **B30** (1984), 3908–3915.
- 11. A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov, *Infinite conformal symmetry in two-dimensional quantum field theory*, Nuclear Physics **B241** (1984), 333–380.
- 12. A. B. Zamolodchikov and V. A. Fateev, *Nonlocal (parafermion) currents in two-dimensional conformal quantum field theory and self-dual critical points in Z_n -symmetric statistical systems*, Soviet Phys. JETP **62** (1985), 215–225.

RIMS, KYOTO UNIVERSITY, KYOTO, JAPAN

Index

- ALEKSANDROV, A. B., 699-I*
 ARAK, T., 994-II
 ARBARELLO, E., 623-I
 AUSLANDER, M., 338-I
 BALABAN, T., 1259-II
 BALLMANN, W., 484-I
 БАШМАКОВА, И. Г., 1612-II
 BEAUVILLE, A., 628-I
 BECK, J., 1400-II
 БЕЛЫЙ, Г. Б., 346-I
 BISMUT, J.-M., 491-I
 BJÖRNER, A., 1408-II
 BLUM, M., 1444-II
 BORHO, W., 350-I
 BOROVoi, M. V., 783-I
 BOS, H. J. M., 1629-II
 BOURGAIN, J., 871-II
 BRANDT, A., 1319-II
 BREZZI, F., 1335-II
 BROUÉ, M., 360-I
 BRYANT, R. L., 505-I
 CARLSSON, G., 574-I
 CATLIN, D., 708-I
 CHANG, S.-Y. A., 715-I
 CHEEGER, J., 515-I
 CHORIN, A. J., 1348-II
 CLEMENS, H., 634-I
 CLOZEL, L., 791-II
 COLIN DE VERDIÈRE, Y., 522-I
 COLLIOT-THÉLÈNE, J.-L., 641-I
 CONNES, A., 879-II
 DAVID, G., 890-II
 DAVIE, A. M., 900-II
 DAVIS, M. H. A., 1000-II
 DE BRANGES, L., 25-I
 DE CONCINI, C., 369-I
 DIPERNA, R. J., 1057-II
 DONALDSON, S. K., 43-I
 DOUADY, A., 724-I
 DRINFEL'D, V. G., 798-I
 ECKMANN, J.-P., 1263-II
 EFFROS, E. G., 906-II
 ELIASBERG, YA. M., 531-I
 EVANS, L. C., 1064-II
 FALTINGS, G., 55-I
 FRANKL, P., 1419-II
 FRENKEL, I. B., 821-I
 GABRIEL, P., 378-I
 GALLAVOTTI, G., 1268-II
 GARNETT, J. B., 917-II
 GAWĘDZKI, K., 1278-II
 GEHRING, F. W., 62-I
 GEMAN, S. AND GRAFFIGNE, C., 1496-II
 GIAQUINTA, M., 1072-II
 ГОДУНОВ, С. К., 1353-I
 GOLDFELD, D., 417-I
 ГОНЧАР, А. А., 739-I
 GRABINER, J. V., 1668-II
 GRIGOR'EV, D. YU., 1452-II
 GINZBURG, V., 840-I
 GROMOV, M., 81-I
 GROSS, B. H., 425-I
 ГЛУСКИН, Е. Д., 924-II
 HARDT, R. M., 540-I
 HAWKINS, T., 1642-II
 HEJHAL, D. A., 1362-II
 HIDA, H., 434-I
 HILDENBRAND, W., 1518-II
 HOLEVO, A. S., 1011-II
 IVRII, V., 1084-II
 IWANIEC, H., 444-I
 JAKOBSON, M. V., 1150-II
 JONES, V. F. R., 939-II
 JOST, J., 1094-II
 KAHANE, J.-P., 1682-II
 KAZHDAN, D., 849-I
 KECHRIS, A. S., 307-I
 KENIG, C. E., 948-II
 KOCH, H. V., 457-I
 KOZLOV, V. V., 1161-II
 KRICHEVSKY, R. E., 1461-II
 KRUIZHILIN, N. G., 749-I
 KRYAZHIMSKII, A. V., 1171-II
 KRYLOV, N. V., 1101-II
 KUNITA, H., 1021-II
 KUPKA, I. A. K., 1180-II
 KUTZKO, P. C., 853-I
 LACHLAN, A. H., 314-I
 LANFORD, O. E., 1385-II
 LEMPET, L., 759-I
 LENSTRA, H. W., 99-I
 LIGGETT, T. M., 1032-II

* I indicates Volume I; II, Volume II.

- MAKAROV, N. G., 766-I
 MANIN, YU. I., 1286-II
 MATHER, J. N., 1190-II
 MEEKS, W. H., 551-I
 MERKURJEV, A. S., 389-I
 MERTENS, J.-F., 1528-II
 MILLER, H., 580-I
 MILMAN, V. D., 961-II
 MIWA, T., 862-I
 MORGAN, J. W., 590-I
 NIKULIN, V. V., 654-I
 ODLYZKO, A. M., 466-I
 ОЛЕБСКИЙ, А. М., 976-II
 ORSZAG, S. A. AND YAKHOT, V., 1395-II
 PAPANICOLAOU, G. C., 1042-II
 PASTUR, L. A., 1296-II
 ПЕРЕТЯТКИН, М. Г., 322-II
 PESIN, YA., 1195-II
 PIPPENGER, N., 1469-II
 POPOV, V. L., 394-I
 QUINN, F., 598-I
 ПАЗБОВ, А. А., 1478-II
 RINZEL, J., 1578-II
 RUH, E. A., 561-I
 SCHEFFER, V., 1110-II
 SCHOEN, R., 121-I
 SCHRIVER, A., 1431-II
 SCHWARTZ, J. T. AND SHARIR, M., 1594-I
 SCHÖNHAGE, A., 131-I
 SEMADENI, Z., 1697-II
 SERIES, C., 1210-II
 SHALEN, P. B., 607-I
 SHAMIR, A., 1488-II
 SHELAH, S., 154-I
 SHOKUROV, V. V., 672-I
 СКОРОХОД, А. В., 163-I
 SMALE, S., 172-I
 SOLONNIKOV, V. A., 1113-II
 SPENCER, T., 1312-II
 STEIN, E. M., 196-I
 STONE, C. J., 1052-II
 SULLIVAN, D., 1216-II
 SUSLIN, A. A., 222-I
 TAKENS, F., 1229-II
 TAUBES, C. H., 1123-II
 TOM DIECK, T., 615-I
 TROMBA, A. J., 1133-II
 URAL'TSEVA, N. N., 1143-II
 VIEHWEG, E., 682-I
 VOGAN, D. A., 245-I
 WEN-TSUN, W., 1657-II
 WENTE, H. C., 565-I
 WILKIE, A. J., 331-I
 WILSON, R. L., 407-I
 WITTEN, E., 267-I
 WOLFF, T. H., 990-II
 WOLPERT, S. A., 777-I
 WÜSTHOLZ, G., 476-I
 ZAGIER, D., 689-I
 ZEHNDER, E., 1237-II
 ZIMMER, R. J., 1247-II